Lisbon School of Economics & Management
Universidade de Lisboa

MESTRADO EM

MÉTODOS QUANTITATIVOS PARA A DECISÃO ECONÓMICA E EMPRESARIAL

TRABALHO FINAL DE MESTRADO

RELATÓRIO DE ESTÁGIO

SPORT LISBOA E BENFICA – ANÁLISE DO PERFIL E COMPORTAMENTO DO SÓCIO

DIOGO MONGE VALENTE

ORIENTAÇÃO:

PROF.ª DR.ª ALEXANDRA MOURA

DR. TIAGO CALDEIRA

OUTUBRO-2023

DOCUMENTO ESPECIALMENTE ELABORADO PARA A OBTENÇÃO DO GRAU DE MESTRADO

Agradecimentos

Começo por agradecer a toda a minha família. Agradecer à minha Mãe e ao meu Pai que sempre me proporcionaram tantos privilégios, especialmente na minha educação, e este mestrado é apenas mais um exemplo disso. Apoiam-me em todos os momentos da minha vida, chamam-me à razão, quando necessário, ajudam-me a levantar quando estou em baixo, e caminham comigo a cada passo. Os meus Tios, Tias, Madrinha e Padrinho, que foram, sem dúvida, as minhas principais referências desde que cheguei à capital e que tanto me têm ajudado neste novo capítulo da minha vida. Os meus Avós, sempre na expectativa de me verem bem, com sucesso profissional, bem cuidado, mas acima de tudo feliz, e que na sua forma muito própria, cuidam de mim, ainda que à distância a maior parte do tempo, e tratam-me como um filho, ou mais que isso. Quero agradecer e dedicar este trabalho ao Manel e à Margarida, para os quais tento, e sempre tentei, ser o melhor modelo que pude. Quero muito que eles sejam bem-sucedidos e felizes, e que saibam que se o primo mais velho conseguiu, eles também conseguirão um dia, e muito mais. Agradeço também aos primos mais velhos, sempre disponíveis, e uma grande ajuda nesta nova fase da minha vida, em Lisboa. Agradecer ao Paulo, que quando estive sobrecarregado com tantas coisas, trouxe gargalhadas e não hesitou em ajudar.

Quero agradecer à professora Alexandra que tanto me ensinou, acompanhou-me e conseguiu sempre puxar por mim da maneira certa, sem deixar que desanimasse. Se não fosse a sua grande ajuda não teria conseguido realizar este trabalho, e muito menos passá-lo para texto. Agradeço também ao Sport Lisboa e Benfica, especialmente a toda a equipa de CRM, claro, pela oportunidade que me foi dada, pela minha primeira experiência profissional, e por descobrir e vivenciar o que é o dia-a-dia num escritório de uma casa tão grande como este clube. Tenho de agradecer especialmente ao Tiago e ao Pedro, que orientaram os meus passos na empresa e ensinaram-me tantas coisas, proporcionando-me 5 meses de uma aprendizagem constante, não só a nível profissional, mas também como pessoa.

Quero agradecer também à família que escolhemos. Todos os amigos que estiverem a meu lado nestes meses e me apoiaram, incentivaram, ou simplesmente acompanharam, agradeço muito a todos eles. Seja em casa, na loja, no estágio, no futebol, estive sempre rodeado de pessoas incríveis que nunca me deram a hipótese de ir-me abaixo.

Por fim, tenho um agradecimento muito especial a fazer, com direito a dedicatória. O Avô, que sempre puxou por mim, que adorava falar do nosso Benfica e que sempre teve aquela sua vontade e gosto em me querer ensinar tantas coisas, acabou por não ver o neto a concluir mais uma etapa. Sei que de toda a família, o Avô foi quem sentiu o maior orgulho ao saber que o neto tinha sido aceite no Benfica. Infelizmente, a poucas semanas de festejarmos este desafio, o Avô apanhou-nos desprevenidos e partiu sem avisar. Obrigado por tudo Avô Monge, esta é para si!

Resumo

Este trabalho tem como principal objetivo estudar o comportamento dos sócios do Sport Lisboa e Benfica e traçar os perfis encontrados, através de uma análise de grupos. O trabalho foi realizado através de um estágio na Direção Comercial e *Marketing* do Sport Lisboa e Benfica, mais concretamente no departamento de CRM. São usados dados de várias bases de dados do SLB, incluindo fontes não só do departamento de CRM, mas também de outros departamentos, como os departamentos de Sócios, *Merchandising*, Bilhética, entre outros. É descrita a forma como os dados foram obtidos, transformados e carregados para uma base de dados, construída no Power BI, de modo a conseguir-se aplicar um modelo capaz de analisar os padrões comportamentais dos sócios. Na descrição do tratamento de dados, são abordados os métodos para imputar *missing values*. Através do R Studio são explorados e testados os algoritmos de agrupamento, juntamente com as diferentes técnicas de agrupamento, tendo sido usado o agrupamento hierárquico. Finalmente, são utilizadas técnicas de visualização, com recurso ao R Studio e Power BI, para estudar o comportamento dos sócios de cada grupo, comparando-os, e destacando os traços que melhor definem os seus perfis. Com esta caracterização dos sócios por grupos é possível, por exemplo, executar campanhas de *Marketing* dirigidas.

Palavras-chave: Análise de grupos; CRM; Comportamento do consumidor; Agrupamento hierárquico; Método de Ward.

Abstract

The main aim of this work is to study the behaviour of Sport Lisboa e Benfica's members, defining customers profiles found, through a cluster analysis. The work was carried out through an internship in the Commercial and *Marketing* Department of Sport Lisboa e Benfica, more specifically in the CRM Department. Data from various SLB databases is used, including sources not only from the CRM Department, but also from other departments, such as Membership, *Merchandising*, Ticketing, among others. It is described how the data was obtained, transformed, and loaded to a dataset, built in Power BI, in order to apply a model capable of analysing members' behavioural patterns. In the description of the data processing, the methods for imputing missing values are discussed. The clustering algorithms are explored and tested using R Studio, along with the different linkage techniques, using hierarchical clustering. Finally, visualization techniques are used, using R Studio and Power BI, to study the members' behaviour of each group, comparing them and highlighting the traits that best define their profiles. With this members' characterisation by groups, it is possible to, for instance, carry out targeted *marketing* campaigns.

Keywords: Cluster analysis; Customer behaviour; Hierarchical clustering; Ward's Method.

Índice

Índice de tabelas

Índice de figuras

1. Introdução

Este trabalho foi realizado no âmbito de um estágio curricular no Sport Lisboa e Benfica, com o objetivo de estudar o comportamento do consumidor, neste caso, dos sócios do clube.

Para realizar este projeto, fui integrado na Direção Comercial e *Marketing* (DCM), mais concretamente na equipa de *Customer Relationship Management* (CRM). A DCM é um dos principais departamentos do Sport Lisboa e Benfica, uma vez que tudo o que é comercializado pelo SLB passa por este departamento, com exceção dos contratos desportivos. Nele estão inseridas diversas equipas que, apesar de desempenharem diferentes funções, têm como objetivo comum assegurar as vendas e expandir o negócio, projetando a marca do clube, e afirmando-o como uma referência na indústria do desporto.

A equipa de CRM é responsável pela análise de dados referentes aos sócios, aplicação de campanhas personalizadas, e comunicação direta com os próprios sócios, através dos canais de comunicação (email ou SMS). Apesar da diversidade de funções desta equipa, o objetivo é sempre o de melhorar a relação entre o sócio e o clube.

Com este contexto presente, o objetivo deste projeto é criar um modelo multivariado para estudar os dados dos sócios e, através de padrões de comportamento, segmentá-los em grupos. No final, o objetivo é realizar uma caracterização dos consumidores através de uma análise de grupos, para promover uma maior personalização, tanto ao nível da comunicação como das campanhas para os diferentes perfis de sócios encontrados.

A análise foi realizada usando uma amostra de 20’000 sócios. Foi utilizado o Power BI para tratar os dados, desenvolver a base de dados e auxiliar na interpretação dos grupos através de *dashboards*. Além do Power BI, o R Studio foi utilizado para tratamento de *missing values* e execução dos algoritmos de segmentação, ou “agrupamento hierárquico” e “agrupamento não hierárquico”, sendo o primeiro a abordagem utilizada, após comparação.

De seguida, é apresentada a forma como a tese está organizada. No Capítulo 2 é descrito ao detalhe o processo de ETL (*extract, transform and load*) dos dados. São exploradas as ferramentas utilizadas durante esta etapa, nomeadamente na obtenção e transformação dos dados, com especial destaque para a ferramenta Power BI, onde foram desenvolvidas as variáveis e construída a base de dados final. São ainda abordados aspetos mais técnicos do Power BI, onde foram resolvidos erros que surgiram durante a construção da base de dados. No Capítulo 3 é feita uma revisão da literatura ao estado da arte e às metodologias mais convencionais utilizadas para a análise ao comportamento do consumidor. Neste Capítulo são ainda aprofundadas as metodologias utilizadas ao longo deste trabalho, assim como as principais razões para a escolha

de umas em detrimento de outras. Foi também explorada a aplicação do Power BI como ferramenta de visualização dos dados, algo fundamental para a discussão e análise dos resultados gerados, pelas metodologias utilizadas. No Capítulo 4 são discutidos os resultados obtidos pelas metodologias abordadas no Capítulo anterior. É feita, inicialmente, uma análise mais geral à amostra e de seguida aos grupos obtidos. Quanto a estes grupos, dois são compostos por sócios menos ligados ao clube, representando a maioria da amostra; outros dois são compostos por sócios com uma forte relação com o clube; e o quinto grupo é formado por sócios com uma ligação relativamente boa, representando um meio termo entre os grupos menos relacionados e os outros dois. Neste Capítulo é feita uma análise mais detalhada aos cinco grupos, assim como às principais diferenças entre eles. No Capítulo 5 é explorado o valor acrescentado por este trabalho, assim como as suas principais aplicações e utilidade para futuros projetos de CRM.

2. Extração, transformação e carregamento dos dados

Neste Capítulo são introduzidos os dados, nomeadamente a sua origem, tratamentos e decisões tomadas para obter a base de dados utilizada no modelo. São abordadas as ferramentas usadas para a extração dos dados e definição das variáveis do modelo, assim como as ferramentas e técnicas utilizadas para transformar, quando necessário, as variáveis e os dados.

2.1. Extração dos dados

Na primeira etapa do estágio, o objetivo foi definir que dados utilizar, ou seja, quais as variáveis em estudo. Assim, foi necessário conhecer melhor as várias áreas do negócio e de que forma poderíamos inseri-las no contexto do modelo. Estas áreas, ou departamentos, que procurei conhecer melhor, referem-se a diferentes tipos de consumos, relevantes para definir perfis na análise ao comportamento do sócio.

Os dados tiveram origem nos seguintes departamentos:

- O departamento de sócios, que providencia informações relativas aos pagamentos de quotas dos sócios, assim como a dados de cadastro referentes a outras matérias no contexto da empresa (por exemplo, a modalidade "Solução Família", que representa vantagens para sócios dentro do mesmo agregado familiar), ou demográficos (por exemplo, a idade e localidade do sócio);
- O departamento de "Mais Vantagens", que procura ampliar e gerir uma rede de parceiros do Sport Lisboa e Benfica, convertendo uma percentagem dos consumos dos sócios, nesses parceiros, em consumo no SLB (através de saldo em carteira virtual). O saldo acumulado fica posteriormente disponível na carteira virtual de cada sócio, para redimir nos consumos realizados em artigos da marca Benfica;
- O departamento de "Bilhética" fornece informações relativas a todas as modalidades de bilhetes comercializadas pela empresa, abrangendo bilhetes não só para jogos de futebol como também para outras modalidades. Também os bilhetes para visitas ao estádio e ao museu estão aqui incluídos. Além da informação diretamente relacionada com bilhetes, este departamento também está encarregue de gerir outros aspetos de bilheteira, como o "Red Pass" (cartão que representa um lugar cativo no estádio, num dado número de jogos, dependendo da tipologia de "Red Pass" adquirido), e o mercado secundário (bilhetes

previamente adquiridos por sócios, que acabam disponibilizados para revenda, dias antes de cada jogo, pela impossibilidade de comparência);

- O departamento de "*Merchandising*" proporciona dados referentes às vendas desta área do negócio, que representam um volume significativo dos consumos dos sócios;
- O próprio departamento de CRM, onde fui integrado, além de utilizar dados referentes aos restantes departamentos já mencionados, também possui informação relativa a campanhas, que é trabalhada unicamente por esta equipa. Estas campanhas têm como objetivo impactar e despertar o interesse dos sócios para diversas vertentes da empresa, visando sempre melhorar a relação entre o sócio e o clube.

Para ter acesso aos dados referentes aos vários departamentos em estudo, foi utilizada a ferramenta "Benficómetro". O "Benficómetro" é uma plataforma utilizada pelas várias equipas, com aplicações diferentes para cada uma delas. No caso da equipa de CRM, o "Benficómetro" representa uma vasta fonte de dados dos sócios, respeitantes ao cadastro e a consumos nas várias áreas do negócio, mencionadas nos departamentos incluídos no estudo. A utilização do "Benficómetro", por parte da equipa de CRM, tem como objetivo auxiliar a execução de várias tarefas, tais como: estudar os dados relativos ao comportamento dos sócios, com o foco nos consumos; elaborar *dashboards* para analisar resultados de campanhas postas em prática, tanto pelo departamento de CRM como por outros; segmentação dos sócios para realizar as comunicações de forma personalizada. O "Benficómetro" permite realizar estas funções porque, além de ser uma plataforma que apresenta estes dados para consulta de todos, também possibilita a extração dos mesmos, através dos seus objetos. Os objetos do "Benficómetro" representam os diferentes tipos de dados que este oferece, existindo um objeto para cada grupo de dados. Assim, com base nas áreas do negócio pretendidas no modelo, os objetos do "Benficómetro" extraídos foram:

- O objeto "Bilhética", com informações relativas a todo o tipo de bilhetes comercializados, tanto de futebol como outras modalidades, e visitas ao estádio e ao museu;
- O objeto "Carteira Virtual", fornecendo dados trabalhados pela equipa de "Mais Vantagens", incluindo saldos, redenções, consumos nos parceiros, etc.;
- O objeto "Red Pass", com os dados referentes aos consumos nesta tipologia de bilhete;
- O objeto "Lista de Espera", com dados referentes aos sócios inscritos na lista de espera para aquisição de "Red Pass", para a época desportiva seguinte;
- O objeto "Encomendas", que representa a generalidade dos consumos realizados pelos sócios;
- O objeto "Produto do Consumo", diretamente relacionado com o objeto anterior, tem dados relativos aos artigos adquiridos, detalhando ainda mais os consumos dos sócios;

- O objeto “Caso”, com informações relativas à abertura de casos, por parte dos sócios, com o intuito de solucionar questões, através do serviço de apoio ao cliente;
- O objeto “Conta”, que é o objeto com mais informação do “Benficómetro”, garantindo, não só os dados demográficos dos sócios, mas também dados referentes a pagamento de quotas, à modalidade “Solução Família”, à “Carteira Virtual”, e ainda à pegada digital do sócio no site oficial do Sport Lisboa e Benfica.

2.2. Construção da base de dados

Após decidir quais os dados a extrair do “Benficómetro”, foi necessário definir as variáveis a implementar no modelo, de forma a elaborar a base de dados. Para tal, foram usados como referência os *dashboards* elaborados pela equipa de CRM. Estes *dashboards* têm como objetivo não apenas analisar o comportamento dos sócios e a sua relação e adesão às campanhas, mas também estudar os resultados das campanhas, individualmente. Os *dashboards* de monitorização de campanhas são constituídos por indicadores que procuram mostrar o impacto que as campanhas tiveram no sócio em termos globais, e também o papel que cada uma desempenhou nos consumos registados. Assim, os *dashboards* gerais apresentam indicadores como: número de campanhas aderidas por sócio; número de consumos total; número de consumos por campanha; etc. Estes indicadores também estão divididos pelas várias áreas do negócio, reunindo a informação de forma isolada para cada área de estudo. Por exemplo, existem indicadores referentes apenas a pagamentos de quotas motivados por campanhas; novas adesões ao programa “Mais Vantagens” também com influência das campanhas; e aquisição de bilhetes a preços inferiores, incentivada por campanhas de descontos. Além destes indicadores mais gerais, foram também tidas em conta outras campanhas que apresentam indicadores que procuram estudar os valores praticados apenas a propósito destas. Algumas destas campanhas foram: Campanha Natal “Mais Vantagens”; Campanha saldo “Mais Vantagens”; Campanha de recuperação de quotas; Campanha de período de descontos; Campanha “Bancada Família”.

Através destes estudos, foi possível inferir quais as variáveis mais apropriadas para incluir no modelo. Inicialmente seriam 192 variáveis, no entanto, devido a limitações na fonte de dados, não foi possível mantê-las a todas. Algumas destas variáveis eram obtidas através de dados com pouca qualidade em “Benficómetro”; outras apresentavam uma quantidade de *missing values*, não substituíveis por zero, muito significativa, pelo que não foi possível encontrar valores que pudessem substituir de forma robusta os valores em falta. Nestas situações, algoritmos de imputação de *missing values* acabam por ser opções pouco confiáveis. Estes algoritmos utilizam os dados existentes para gerar os que faltam, no entanto, quando a quantidade de *missing values* é muito significativa, o algoritmo dispõe de pouca informação para desenvolver os valores

pretendidos. Em alguns casos, as variáveis desejadas eram demasiado complexas de desenvolver, principalmente pela forma como a informação está estruturada no "Benficómetro". Assim, após retirar do modelo as variáveis que apresentavam estas limitações, obtivemos uma base de dados com 133 variáveis. Foi obtida uma listagem destas variáveis e elaborada uma tabela em Excel para organizá-las e detalhá-las. Nesta tabela, cada linha representa uma variável e cada coluna uma característica. A tabela é composta por 133 linhas, que correspondem às variáveis definidas, e pelas seguintes 7 colunas:

- Nome, definindo de forma simples o que a variável pretende explorar;
- Categoria, separando as variáveis pelas diferentes áreas de estudo do modelo. As categorias presentes na tabela são: Bilhética; Campanhas; Demográfica; Digital; Encomendas; Mais Vantagens; Referência; Registo de Sócio; Restrição. Estas categorias já foram contextualizadas, com exceção das categorias Referência e Restrição. As variáveis de referência servem de indicadores-chave para avaliar aspetos mais gerais do perfil do sócio, por exemplo, "Maior/menor de idade" ou "Já adquiriu produtos *merchandising*". Quanto às variáveis de restrição, são utilizadas para auxiliar a definição da amostra, por exemplo, "Sócio pagante" ou "Máxima antiguidade da dívida";
- Subcategoria, que é utilizada principalmente para distinguir aspetos do negócio inseridos em categorias muito abrangentes. As subcategorias presentes na tabela são: Bancada Família; Definição da Amostra; Geral; Horizonte Temporal; Mercado Secundário; *Merchandising*; Modalidades; Partilha; Período de Descontos; Quotas; Red Pass; Solução Família; Transações Gerais; Visitas estádio/museu. A maior parte destas categorias já está contextualizada, com exceção das "Bancada Família" e "Período de Descontos" que representam variáveis referentes a estas campanhas em concreto, "Definição da Amostra" e "Horizonte Temporal" que são auxiliares para definir as restrições e o horizonte temporal da amostra, "Geral" que contém variáveis relacionadas com aspetos mais gerais do registo do sócio (por exemplo, "Categoria do sócio" ou "Última entrada no estádio"), e "Partilha", onde as variáveis são referentes à modalidade de partilha de bilhetes entre sócios;
- Classificação, que distingue as variáveis entre numéricas e categóricas;
- Acessibilidade, que é utilizado como atributo auxiliar para o desenvolvimento das variáveis. Se a variável está exibida de forma direta no "Benficómetro", a sua acessibilidade é "Direta", uma vez que o dado está disponível para ser consultado. Se, por outro lado, for necessário recorrer a técnicas de programação, como aplicação de filtros e condições, para chegar ao dado pretendido, a variável é considerada "Não Direta" quanto à sua acessibilidade;
- Descrição, que tem como objetivo explicar em que consiste a variável, de forma sucinta;

- Procedimento, que pretende relatar de que forma foi alcançado o dado que a variável procura estudar. Dependendo da acessibilidade, o procedimento pode ser simples, nos casos em que o dado já está representado de forma direta no "Benficómetro", ou mais trabalhoso, quando é necessário proceder à utilização de técnicas de programação, deixando relatado, de forma concisa, o procedimento adotado.

Na Tabela 1 estão exemplificadas duas linhas da tabela de variáveis, representando duas variáveis.

Tabela 1 – Exemplificação da tabela de variáveis

Nome	Categoria	Subcategoria	Classificação	Acessibilidade	Descrição	Procedimento
Sócio Pagante	Restrição	Definição da Amostra	Categórica	Direta	Define se a modalidade do sócio implica pagamentos de quotas ou se se trata de um sócio isento	Coluna "Socio_Pagante__c" direta, no objeto da conta, filtrada com o valor "True" para a definição da amostra
Última entrada no estádio	Registo de Sócio	Geral	Numérica	Não Direta	Tempo decorrido desde a última entrada do sócio no estádio, em meses	Devolve a diferença entre a data da última entrada no estádio e a data de hoje, em meses

Quanto à definição da amostra, foi decidido que, apesar de trabalharmos com os dados de todos os sócios disponíveis em "Benficómetro", o modelo seria apenas aplicado a uma amostra da população de sócios. Esta decisão foi tomada com base na dimensão do modelo, devido ao elevado número de variáveis definidas, 133, e, simultaneamente, devido às limitações de capacidade e memória da ferramenta utilizada, R Studio, para realizar, posteriormente, o "clustering". A base de dados foi testada no R, utilizando amostras de diferentes dimensões, e chegámos à conclusão que uma amostra de 20'000 sócios representa o número máximo com que se conseguiu trabalhar com o R, devido às questões de armazenamento e memória. Assim, foi selecionada uma amostra de 20'000 sócios, com a particularidade de esta ser uma amostra pseudo-aleatória e com um horizonte temporal de duas épocas desportivas. A questão da pseudo-aleatoriedade vem da forma como foram selecionados os sócios pertencentes à amostra, uma vez que estes devem respeitar três critérios, devido a restrições do modelo:

- A tipologia do consumidor deve ser "sócio". No "Benficómetro" também estão presentes dados referentes aos adeptos (simpatizantes do clube que não são sócios), porém estes

dados carecem de qualidade e quantidade, pelo que poderiam gerar resultados pouco precisos e conclusões com pouco valor;

- O sócio deve ser pagante. Esta restrição tem o objetivo de excluir os sócios que não realizam pagamentos de quotas, uma vez que este é um dos consumos analisados, evitando assim a existência de sócios isentos, por sua vez com dados incompletos neste aspeto do negócio;
- A máxima antiguidade da dívida deve ser igual ou inferior a 12 meses. Este foi um indicador definido especificamente, pela equipa de CRM, para este modelo, restringindo os sócios quanto aos seus pagamentos de quotas. O objetivo é limitar as dívidas de quotas até um máximo de 12 meses, uma vez que sócios com uma dívida superior a esta são considerados sócios em *churn*, ou seja, sócios perdidos (ou que serão perdidos num futuro próximo) ou difíceis de recuperar. Este tipo de sócio apresenta, logicamente, uma relação mais distante com o clube, o que se traduz em lacunas na matéria de consumos, ou seja, fraca qualidade dos dados. Com esta restrição é possível incluir apenas os sócios com as quotas em dia, pagamentos adiantados, e dívidas de, no máximo, 12 meses, evitando assim a presença dos sócios mais alheios ao clube.

Quanto ao horizonte temporal, foi definido que seriam incluídas as épocas 2021/2022, começando a 1 de julho de 2021 e terminando a 30 de junho de 2022, e 2022/2023, começando a 1 de julho de 2022 e terminando a 30 de junho de 2023. É de realçar que a definição do horizonte temporal impacta apenas os dados relacionados com consumos, uma vez que outro tipo de dados, nomeadamente demográficos (por exemplo, idade ou localidade), não são influenciados pelo período de tempo definido. Outro motivo pelo qual foram escolhidas apenas estas duas épocas desportivas foi o impacto do período pandémico nas épocas anteriores. Apesar de na primeira metade da época 2021/2022 ainda existirem algumas limitações devido à pandemia, estas foram pouco impactantes, pelo que considerámos que seria apropriado incluí-la também no horizonte temporal. Foram ainda definidas três restrições, presentes na tabela das variáveis, para definição do horizonte temporal: Época Bilhética (auxiliar); Época Carteira Virtual (auxiliar); Época Encomendas (auxiliar). Estas variáveis são auxiliares e pretendem corrigir erros entre a associação das datas dos consumos com a época desportiva, nos três objetos mencionados. Desta forma, asseguramos que consumos com datas entre 1 de julho de 2021 e 30 de junho de 2022 pertencem à época desportiva 2021/2022, e, da mesma maneira, consumos com datas entre 1 de julho de 2022 e 30 de junho de 2023 correspondem à época desportiva 2022/2023.

2.3. Power BI

Após definir a base de dados do modelo e o horizonte temporal a estudar, o objetivo foi implementar as variáveis através da ferramenta Power BI. Para tal, é necessário começar por importar os dados pretendidos do "Benficómetro" para o Power BI. Foram escolhidos os objetos do "Benficómetro" mencionados anteriormente, de forma a dispor de todos os dados necessários para elaborar as variáveis definidas.

Depois de realizar a importação, é possível observar, através de uma interface no Power BI, o esquema do modelo, onde os objetos estão dispostos de forma isolada, sem quaisquer ligações entre eles. Para desenvolver as variáveis pretendidas é necessário criar ligações, manualmente, através de uma chave em comum entre os vários objetos. De modo a proteger e salvaguardar os dados pessoais dos sócios, é utilizada uma chave fictícia composta por uma combinação de números e letras à qual se dá o nome de "Id". O "Id" é único para cada sócio e permite identificá-lo sem que seja necessária informação mais pessoal, como, por exemplo, o nome próprio. O "Id" está presente na "Conta" e nos consumos (individualmente) de todos os objetos, tornando possível associar estes consumos à conta do respetivo sócio. Por exemplo, no caso do objeto "Bilhética", cada linha corresponde a um bilhete, e em cada uma destas linhas, além de todos os dados referentes ao bilhete, existe uma coluna chamada "Conta__c", preenchida com o "Id" do sócio que adquiriu o bilhete em causa. Este processo é articulado para os restantes objetos, relacionando cada linha de consumo ao sócio correspondente.

Além dos objetos do "Benficómetro", foi importada outra fonte de dados, à qual se deu o nome "Anos Red Pass", necessária para desenvolver algumas variáveis. Esta base de dados consiste numa tabela com duas colunas: o "Id" do sócio, e o número de épocas desportivas em que o sócio deteve "Red Pass". Este dado representa uma lacuna no "Benficómetro", uma vez que não está disponível na plataforma, tendo sido necessário recorrer a esta base de dados adicional. À semelhança dos objetos do "Benficómetro", esta base de dados é apresentada na forma de uma tabela, ligada ao objeto da "Conta" através do valor do "Id".

Assim, o objeto "Conta" estabelece ligações com os seguintes objetos: Bilhética; Red Pass; Anos Red Pass; Lista de Espera; Carteira Virtual; Encomendas; Caso. Além destes objetos, existe também uma ligação entre os objetos "Encomendas" e "Produto do Consumo", através de uma chave semelhante ao "Id", que em vez de identificar o sócio, associa um dado produto ao consumo total onde pertence.

Feitas as ligações entre objetos, o passo seguinte foi o desenvolvimento das variáveis. Para implementar as variáveis na base de dados, utilizámos a interface do Power BI "Vista de tabela", onde é possível visualizar todos os objetos no formato de tabelas. As variáveis devem ser desenvolvidas na tabela referente ao objeto "Conta", salvo raras exceções, uma vez que esta é a

tabela principal, onde estão conectados todos os outros objetos, fora o "Produto do Consumo". Assim, é possível adicionar novas colunas à tabela (que correspondem às variáveis) e utilizar a informação disponível em todos os objetos para alcançar os dados pretendidos, executando as variáveis definidas.

Inicialmente, implementamos as variáveis de acessibilidade direta. Estas são variáveis que já estão incluídas na tabela da "Conta", pelo que é apenas necessário confirmar os valores que são apresentados, por uma questão de controlo de qualidade dos dados. Neste grupo de variáveis existem duas exceções: Tipologia; Sócio Pagante. Uma vez que estas variáveis fazem parte do conjunto de restrições para definição da amostra, é necessário aplicar os filtros: "Sócio" na variável Tipologia; "True" na variável Sócio Pagante. Estes filtros foram aplicados na *Query* do Power BI, uma interface que permite carregar, restruturar, moldar e combinar bases de dados, aplicando essas alterações ao nível do Power BI *desktop*, interface utilizada para desenvolver toda a base de dados e, posteriormente, analisá-la graficamente.

Após a verificação das variáveis diretas, o objetivo foi desenvolver as restantes, não diretas, através de técnicas de programação como filtros, condições, funções ou até variáveis auxiliares. Por exemplo, a variável "Número de Encomendas (Quotas)" foi desenvolvida através do uso de filtros e funções. Para tal, é feito o cálculo (função *calculate*) da contagem distinta (função *distinctcount*) de encomendas, aplicando o filtro de correspondência entre o "Id" de sócio, associado ao consumo e ao objeto "Conta" do sócio, e o filtro com a tipologia da encomenda, que tem de conter a palavra "Quota" (função *containstring*). No caso da variável "Máxima antiguidade da dívida", foi utilizada uma condição com o auxílio da variável "Antiguidade da dívida, em meses". Se esta variável tomar um valor igual ou inferior a 12, obtemos o valor "*True*", caso contrário "*False*".

Além das técnicas utilizadas, também foi necessário adaptar algumas variáveis, nomeadamente, pela forma como o "Benficómetro" apresenta os seus valores. O caso mais comum é a tradução de variáveis categóricas, uma vez que o "Benficómetro", por defeito, apresenta algumas variáveis com valores num formato numérico em vez de nominal, o que dificulta a sua compreensão e análise. Assim, com recurso ao "Benficómetro" e a variáveis auxiliares ou condições, foi possível legendar os valores destas variáveis. Por exemplo, na variável "Categoria do sócio" foi realizado o processo de tradução, uma vez que esta variável é apresentada na forma numérica em "Benficómetro". Se o valor em "Benficómetro" for:

- 19, o sócio é "Infantil";
- 20, o sócio é "Correspondente";
- 26, o sócio é "Maior";
- 28, o sócio é "Juvenil";

- 29, o sócio é “Reformado”;
- 42, o sócio é “Sub-23”;
- Os restantes valores serão traduzidos como sócio “Família”.

Note-se que as 35 variáveis categóricas não foram usadas para executar o algoritmo da análise de grupos, mas sim para auxiliar a análise ao perfil dos sócios de cada grupo, juntamente com as restantes 98 variáveis. Para o algoritmo de agrupamento foram utilizadas somente as 98 variáveis numéricas.

Através da utilização e combinação destas técnicas foi possível desenvolver, como mencionado acima, as 133 variáveis propostas para a base de dados, seguindo-se o tratamento das mesmas.

2.4. Revisão e resolução de erros da base de dados

Após desenvolver as variáveis através do Power BI, foi necessário realizar o tratamento dos dados, através da revisão dos valores gerados, variável a variável. Este tratamento de dados foi composto por duas etapas: controlo de qualidade; inspeção e tratamento de *missing values*.

Na primeira, o objetivo foi realizar um controlo de qualidade para garantir que os dados oferecem valores coerentes com a informação disponível em “Benficómetro”, mas também perceber se existem valores não expectáveis, por exemplo taxas superiores a 100% ou variáveis respeitantes a consumos com valores negativos. Estes erros podem ter origem em informação de má qualidade ou mal estruturada em “Benficómetro”, ou podem ser erros de código durante a etapa anterior. Algumas das variáveis apresentavam resultados indesejáveis numa quantidade considerável de observações, e não sendo possível corrigi-las em Power BI, optou-se pela sua remoção do modelo, uma vez que a sua permanência na base de dados poderia gerar erros, maior complexidade do modelo e resultados de pior qualidade. As variáveis que apresentavam erros de código, sem qualquer problema na qualidade dos dados, foram corrigidas ao nível do Power BI, com recurso às técnicas de programação já mencionadas. Quanto às restantes variáveis, foram também revistas e, não tendo qualquer erro ou valor não expectável, mantidas no modelo sem realizar qualquer alteração.

Na segunda etapa do tratamento da base de dados, procedeu-se a uma análise e tratamento de *missing values*.

A base de dados foi carregada para o R Studio, onde foi possível visualizar, na forma de tabela, o número de *missing values* para cada variável, permitindo saber quais as variáveis a necessitar de tratamento. Os *missing values* podem ser provenientes de erros de código, durante o tratamento e construção das variáveis, ou erros e limitações da fonte dos dados, neste caso, do “Benficómetro”.

Após uma primeira observação desta tabela para entender a dimensão dos *missing values*, em cada uma das variáveis, foram averiguadas as possíveis origens destes valores e de que forma podem ser corrigidos, de acordo com as razões seguintes:

- Os *missing values* correspondem a zero. Esta foi a origem de *missing values* mais frequente, uma vez que, por defeito, nas tabelas de dados são apresentadas células em branco ou com o valor "NA" em vez de "0". Para resolver esta questão, basta substituir as células em branco ou com o valor "NA" por "0", na tabela da "Conta", onde foram desenvolvidas as variáveis do modelo, ou no ficheiro que contenha a base de dados final;
- Erros no "Benficómetro". Algumas das variáveis apresentavam *missing values* devido a problemas com o "Benficómetro", nomeadamente, dados a aguardar atualizações, ou erros na caracterização de alguns consumos. Estas situações eram muito mais pontuais que os casos anteriores, e, sabendo os verdadeiros valores em falta, optou-se por realizar uma imputação manual com os dados corretos;
- Os *missing values* correspondem a valores diferentes de zero. Estes foram os casos mais difíceis de solucionar, uma vez que estes *missing values* não poderiam ser substituídos por 0, como no primeiro caso, mas também não estava disponível a informação real em "Benficómetro", como na situação anterior. Estas foram as situações mais escassas, tendo-se verificado apenas nas duas variáveis "Longevidade de adesão às mais vantagens, em meses" e "Longevidade do registo do sócio no site (meses)". O erro está relacionado com o facto de alguns sócios serem aderentes "Mais Vantagens" e estarem inscritos no site, não havendo, porém, dados quanto às datas de inscrição e adesão a estas modalidades. Ou seja, não é um erro do próprio "Benficómetro", mas sim uma falha de registo. No caso da variável "Longevidade de adesão às mais vantagens, em meses", foi corrigido o valor para os dois sócios em questão, associando a data de adesão à data do primeiro consumo com a utilização da "Carteira Virtual". Já na variável "Longevidade do registo do sócio no site (meses)", foram reportados 38 sócios inscritos no site, a maior parte com apenas uma visita, mas sem data de inscrição. Assim, e na impossibilidade de chegar ao valor correto, foi imputado manualmente o valor "12" (meses), que corresponde, aproximadamente, ao valor médio desta variável, na amostra de 20'000 sócios.

Após corrigir os erros e imputar os *missing values*, a base de dados ficou pronta para iniciarmos a aplicação e testagem dos algoritmos de segmentação, no R Studio.

3. Metodologias

Neste Capítulo é feita uma revisão da literatura relativamente à utilização das grandes bases de dados para realizar análises ao comportamento do consumidor e sobre qual o papel desempenhado pelo CRM nesta temática, procurando conhecer melhor o estado da arte. De seguida, serão exploradas as metodologias para executar a análise de grupos, com foco na abordagem selecionada neste estudo. Finalmente, abordamos a forma como utilizámos o Power BI para visualizar e estudar os resultados, através dos *dashboards* que desenvolvemos.

3.1. Revisão de literatura para análise ao comportamento do consumidor

Antes dos dados tomarem a importância que desempenham nos dias de hoje, estes eram recolhidos apenas para registo das empresas, tendo utilidades muito limitadas e específicas. Inicialmente, o registo dos dados não era considerado um ativo de valor das empresas, e eram apenas utilizados para cálculos de contabilidade ou de receitas publicitárias. Com o surgimento da "*Big Data*" o propósito destes dados mudou, passando a ser visto como um ativo valioso, a explorar por todo o tipo de empresas (Anshari et al., 2019). A chegada da "*Big Data*" foi incitada pela globalização da Internet, acessível a todos, promovendo o aparecimento de grandes quantidades de dados, de diversas fontes, tornando o sucesso das empresas dependente da captação e boa utilização destes dados (Li et al., 2019). Além dos métodos pouco desenvolvidos, na época, os dados recolhidos eram mal estruturados e em grandes escalas, tornando o seu processamento e tratamento exaustivos e pouco eficientes, porém, necessários para a evolução das empresas (Khade, 2016). Segundo, Khade (2016, p. 986), os grandes desafios, para a época, no processamento das grandes bases de dados seriam a otimização do "armazenamento, pesquisa, distribuição, transferência, análise e visualização" dos dados. Assim, a importante competência para extrair os dados e trabalhar com eles, traduz-se num forte indicador do desenvolvimento das empresas e da sua capacidade competitiva, aplicável a várias indústrias (Anshari et al., 2019; Li et al., 2019). A "*Big Data*" é também caracterizada, ao longo da literatura, pelos 5 V's: o seu volume, pela dimensão das bases de dados muito superior às mais convencionais; a velocidade com que são, continuamente, geradas estas bases de dados; a variedade de formatos com que as bases de dados podem ser apresentadas no momento da extração; a variabilidade de fontes de dados que podem ser utilizadas para gerar as bases de dados e a informação pretendida; e a volatilidade ou veracidade, uma vez que também é possível extrair bases de dados de fraca qualidade (Anshari et al., 2019; Hofacker et al., 2016; Miah et al., 2017).

O grande volume de dados gerado, pode beneficiar diversas áreas como a ciência, saúde, indústria financeira, segurança, governos, e até sociedades, mas também, o setor empresarial (Anshari et al., 2019), onde está introduzido este estudo. Todas estas indústrias valorizam a recolha dos dados, reconhecendo o seu valor e as vantagens competitivas associadas (Provost & Fawcett, 2013). Como a extração pode gerar tanto bases de dados estruturadas como não estruturadas, foi necessário implementar a análise de dados nestas indústrias, de modo a dinamizar o *modus operandi* dos dados, contribuindo para melhores decisões empresariais (Surendro, 2019). Segundo o autor (Nauck, 2013, p. 2), "a Big Data sem análise, é apenas uma grande quantidade de dados" e assim, com a análise de dados, o objetivo é revolucionar a maneira como estes são utilizados.

Uma das áreas em estudo que alterou drasticamente o modo de operar com a chegada da "Big Data" foi o *marketing*. Segundo vários autores, o *marketing* e a tomada de decisão no setor empresarial, passaram a ser fortemente influenciados pela análise de dados dos consumidores, o que revolucionou a forma como o *marketing* das empresas é gerido (Erevelles, 2015; Hofacker et al., 2016). Assim, a análise de dados passa a ter um papel fundamental nas empresas. No contexto do *marketing* e deste trabalho, focamos a pesquisa num tipo concreto de análise de dados: a análise ao comportamento do consumidor. Esta análise tem o objetivo de atribuir valor prático aos dados disponíveis sobre os consumidores, melhorando vários aspetos do negócio, nomeadamente o aumento das vendas, a relação com o consumidor, e o *marketing* da empresa, através do conhecimento mais apurado do seu público-alvo (Anshari et al., 2019; Khade, 2016). Para desempenhar esta análise, as empresas concentram-se em melhorar a recolha de dados respeitantes aos seus consumidores e aos produtos que adquirem, para que, deste modo, obtenham um conhecimento mais detalhado quer das compras realizadas pelos consumidores, quer dos produtos comercializados pela empresa (Anshari et al., 2019). Uma das formas encontradas para melhorar e expandir esta recolha de dados é a utilização dos cartões de cliente ou da adesão a sócios, como é o caso do Sport Lisboa e Benfica, de modo a facilitar a monitorização e obtenção de dados dos consumidores (Kanavos et al., 2018). Os dados referentes ao comportamento do consumidor são muito abrangentes, gerando bases de dados tanto mais extensas quanto maior for o negócio e melhor a recolha dos dados. É importante mencionar que estes dados têm também um propósito preditivo, uma vez que estudar o comportamento passado dos consumidores, ajuda a prever como será no futuro (Leventhal, 2018). Ainda assim, estudar o comportamento do consumidor não garante sempre resultados preditivos precisos, uma vez que o seu comportamento pode rapidamente mudar face a novas necessidades ou alterações. Por isso, o objetivo é estudar o comportamento do consumidor, procurando seguir padrões comportamentais, obtendo assim previsões mais exatas e detetando, mais facilmente, alterações de comportamento (Surendro, 2019).

Com as irregulares alterações do mercado e do comportamento do consumidor, o objetivo das empresas é estarem o mais atualizadas possível, garantindo a rentabilidade do negócio. Desta forma, muitas empresas asseguraram a criação de um departamento de *Customer Relationship Management* (CRM) que deve prever o comportamento do consumidor, através do histórico e dados disponíveis, aumentando a satisfação dos consumidores e a sua ligação à empresa (Abirami & Pattabiraman, 2016). A equipa de CRM de uma empresa é também responsável por estabelecer, consolidar e desenvolver as relações entre os consumidores e a empresa (Berry et al., 1983; Chan et al., 2018). Estas relações desenvolvidas com os consumidores são geridas pelas equipas de CRM e não são vistas apenas com uma perspetiva de negócio, mas sim como uma aproximação entre consumidores e empresa, criando laços pessoais entre os envolvidos (Abdullah Al-Suraihi et al., 2020). Antes do CRM concentrar as suas atividades em torno dos clientes, era visto como uma ferramenta para utilizar no setor das tecnologias de informação, em meados da década de 1990. Só no século XXI, com a já referida globalização da Internet, o CRM passou a ser visto como uma ferramenta ou tecnologia na gestão das empresas, para gerir e acompanhar as interações com os consumidores, mas também, melhorar a eficiência operacional da empresa (Abdullah Al-Suraihi et al., 2020; Ahmed & Aissa, 2018). O CRM pode ser definido através de vários propósitos com que é utilizado: pode ser visto como a filosofia da empresa, aplicando as bases e valores à abordagem da empresa para com os consumidores; pode ser uma ferramenta de estratégia empresarial, oferecendo serviços mais apelativos e personalizados, aumentando a satisfação do consumidor e a sua lealdade para com a empresa; é também definido como uma parte ativa no processo empresarial, através da expansão do conhecimento sobre o consumidor; e ainda, do ponto de vista tecnológico, assegurando relações entre consumidores e empresa mais estimuladas e vincadas (Ahmed & Aissa, 2018).

É também de realçar a especial responsabilidade do departamento de CRM pelas campanhas de *marketing*, tanto na criação das mesmas como na comunicação aos consumidores. Um dos principais problemas encontrados nesta temática é a pouca eficácia nas estratégias de *marketing*, uma vez que muitas empresas, principalmente aquelas cujo CRM ainda não é devidamente valorizado, simplificam as estratégias de *marketing* resultando numa dificuldade de gerir e prolongar a relação com os consumidores. Nestas situações, as empresas têm o seu foco na relação a curto prazo com os consumidores, algo que o CRM pretende corrigir, personalizando e adaptando as estratégias de *marketing* para os seus atuais e potenciais consumidores. O papel da "Big Data" nestes contextos é auxiliar e servir como principal fonte a personalização conseguida pelo departamento de CRM, de modo a melhorar estas relações e alcançar um maior número de consumidores, da maneira correta (Anshari et al., 2019).

Finalmente, um tema muito abordado pelas equipas de CRM é a dificuldade em reter consumidores propícios a abandonar a sua ligação com a empresa (consumidores em "churn"). O

termo “churn” foi o nome dado aos consumidores dentro deste espetro, e a sua gestão é vista como uma das funções mais desafiantes das equipas de CRM. A principal medida para segurar este tipo de consumidor é direcioná-los a campanhas de *marketing* específicas, altamente personalizadas, para os seus contextos (Hadden et al., 2007). A retenção de consumidores é um dos aspetos que define a competitividade da empresa, e é vista como uma das principais estratégias de *marketing*, adotadas pelos departamentos de CRM (Kim et al., 2004). Este problema é uma realidade com que a equipa de CRM do Sport Lisboa e Benfica tem de lidar diariamente, uma vez que, neste tipo de empresas, onde devem ser realizados pagamentos de quotas, os casos em que os sócios param de pagar quotas e, consequentemente, acumulam dívidas altas, é muito comum, aumentando significativamente o número de sócios em “churn”. Este projeto também tem em conta este contexto e um dos objetivos é conseguir fazer uma distinção entre os sócios menos ligados ao clube, personalizando cada vez mais as campanhas, para os diferentes tipos de sócios em “churn”.

3.2. Metodologias para análise de grupos

Na Secção anterior vemos a importância da análise de dados no conhecimento das empresas sobre os seus consumidores, traduzindo-se num forte indicador da capacidade competitiva na indústria.

Os comportamentos são estudados através de análises a bases de dados recolhidas pelas empresas, classificando os consumidores, e agrupando-os, de acordo com as suas semelhanças (Abdi & Abolmakarem, 2019; Kanavos et al., 2018). Existem vários tipos de modelos utilizados para explorar os dados, não só, relativos ao estudo do comportamento do consumidor, como também a outras bases de dados. Existem modelos de classificação, agrupamento e associação, sendo os de classificação mais utilizados como modelos de estimação, enquanto os restantes são considerados modelos de identificação, onde são procurados comportamentos e evidências que suportem a tomada de decisões. Assim, no contexto deste estudo, o mais apropriado é realizar a análise através de modelos de identificação, mais concretamente, de agrupamentos de consumidores (Aggarwal & Yu, 1999; Kaya Gülağız & Şahin, 2017; Savaş et al., 2012). Através destas análises, os consumidores são segmentados em grupos, onde cada grupo tem um perfil de sócio designado. Este procedimento de categorização de consumidores é considerado uma estratégia de *marketing* muito útil na gestão das relações entre a empresa e os consumidores, uma vez que, feita a análise, é possível conhecer as suas características, comportamentos e necessidades, consoante o grupo a que pertencem (Abdi & Abolmakarem, 2019; Bose & Chen, 2015; Hsu et al., 2012). As segmentações dos clientes são baseadas nos atributos demográficos (como é o caso do género e idade), geográficos (nomeadamente o país e cidade onde residem), psicográficos (que procuram explorar o estilo de vida e de que forma a empresa está inserida nele,

pelo que são dados muito concretos e próprios de cada empresa) e comportamentais, ou hábitos consumistas que, à semelhança dos psicográficos, podem variar bastante entre empresas (por exemplo, no caso do Sport Lisboa e Benfica, "valor gasto para bilhetes em jogos de futebol" não é um dado aplicável a uma empresa focada em têxteis). Conjugando as diferentes tipologias de dados que as empresas recolhem e estudam, é possível compreender o comportamento dos consumidores, implementando políticas e campanhas cada vez mais personalizadas para os vários grupos obtidos (Abdi & Abolmakarem, 2019; Bose & Chen, 2015). Desta forma, as empresas conseguem segmentar o mercado, adotando abordagens e estratégias que correspondam às necessidades dos consumidores de uma maneira mais precisa e personalizada (Mohammadi Nasrabadi et al., 2013).

A este processo de segmentação é também dado o nome de "clustering". A análise de grupos, ou análise de "clusters", consiste na segmentação do objeto de investigação (neste caso, os sócios do Sport Lisboa e Benfica), com base na proximidade entre as características das observações, evidenciando, principalmente, as diferenças no seu comportamento enquanto consumidores. Os grupos formados são também designados por "clusters", e devem ter uma elevada similaridade intra grupos e baixa similaridade entre grupos, ou seja, as observações dentro de um mesmo grupo devem ser semelhantes, mas os grupos devem apresentar diferenças evidentes entre si (Abirami & Pattabiraman, 2016; Bair, 2013; Li et al., 2019). Para este estudo foram utilizadas técnicas de agrupamento (algoritmos de "clustering") que combinadas com técnicas de visualização, nomeadamente através do Power BI, proporcionam uma análise mais intuitiva e completa dos dados.

Existem muitos algoritmos de "clustering" que, geralmente, produzem resultados diferentes para uma dada base de dados. A qualidade da análise de grupos depende fortemente do algoritmo utilizado, que deve ser escolhido com base na sua adequabilidade face à base de dados em estudo, ou a objetivos em concreto do projeto. Muitas vezes, as decisões quanto ao algoritmo são pouco ponderadas ou muito limitadas devido a custos de execução ou dos próprios *softwares*, requisitos de memória muito exigentes, ou carência de informação e pesquisa por parte dos utilizadores (Ackerman & Ben-David, 2016). Por isso, procedemos a uma análise às principais abordagens de "clustering" assim como aos métodos que definem a forma como estas abordagens geram os resultados.

Como já abordado, os modelos de "clustering" permitem segmentar os consumidores em vários grupos de acordo com as semelhanças entre eles. Estas semelhanças são calculadas com base nas características dos consumidores, através da medição das distâncias dos valores numéricos. Os valores numéricos apresentados na base de dados podem ser dispostos na forma de matriz, em que neste caso cada linha é um sócio e cada coluna um valor numérico, correspondente a uma

variável. As linhas da matriz podem ser vistas como vetores num espaço multidimensional, sendo a dimensionalidade da matriz correspondente ao número de variáveis contidas na base de dados. Através dos vetores (representações dos sócios) são calculadas as distâncias entre eles por medidas de dissemelhança. As medidas mais convencionais são a Distância Euclidiana, a Distância de Manhattan e a Distância de Minkowski. Estas distâncias medem a dissemelhança entre dois pontos, ou seja, as observações são mais semelhantes quanto menor o valor da distância entre elas (Bezdek et al., 2007; Kaya Gülağız & Şahin, 2017; Murtagh & Contreras, 2012). Neste estudo utilizámos a distância euclidiana, que representa a distância entre dois pontos, no espaço multidimensional, calculada através dos vetores, construídos pela matriz inicial. O cálculo da distância euclidiana é feito através da raiz quadrada da soma dos quadrados das diferenças, como pode observar-se na fórmula seguinte:

$$D_{ij} = \sqrt{\left(x_{i1} - x_{j1}\right)^2 + \cdots + \left(x_{ip} - x_{jp}\right)^2} = \sqrt{\sum_{k=1}^{p} \left(x_{ik} - x_{jk}\right)^2}$$

- D_{ij} representa a distância euclidiana entre os sócios *i* e *j*;
- p é o número de dimensões no espaço vetorial, ou seja, o número de variáveis (colunas) da base de dados;
- x_{ik} e x_{jk} são os valores referentes à k-ésima variável dos sócios *i* e *j*.

Calculando as distâncias possíveis entre cada par de observações, é gerada uma matriz com todos estes valores, designada matriz de dissemelhança (Jobson, 2012; Posit team, 2023; Sharma, 1995).

Após escolher a distância para calcular as dissemelhanças, formando a matriz de dissemelhança, é necessário escolher a técnica de agrupamento mais apropriada, de acordo com os objetivos do estudo. A escolha é feita entre duas categorias principais de agrupamento: o "cluster" hierárquico e o "cluster" não hierárquico. Estas foram as técnicas abordadas neste projeto, uma vez que, são as mais referenciadas em matérias de análise de grupos para o estudo do comportamento do consumidor.

Quanto aos algoritmos de "clustering" não hierárquicos, estes agrupam as observações de uma forma mais direta, comparativamente aos hierárquicos, sem necessitarem de muitas iterações, e, em consequência, não requerem softwares com grande capacidade de memória. Nos algoritmos de agrupamento não hierárquico é necessário definir antecipadamente o número de "clusters" desejados, uma vez que, é a partir deste número que o algoritmo vai definir o centro dos grupos. Tome-se o exemplo do algoritmo não hierárquico com o método K-Means, como técnica de agrupamento, bastante utilizado e referenciado pela literatura. Seja k o número de "clusters" definido e n o número de observações da amostra, o algoritmo começa por selecionar k

observações para gerar os k grupos. As restantes $(n - k)$ unidades amostrais são distribuídas pelos grupos mais próximos, através da medida de dissemelhança elegida. O centroide de cada "cluster" é calculado para, posteriormente, realocar as observações de cada grupo para o "cluster" cujo centroide se encontra mais próximo da unidade amostral a realocar. Este procedimento repete-se até atingir um determinado critério de convergência, definido no método de agrupamento (existem vários métodos de agrupamento com diferentes critérios de convergência que não serão explorados neste estudo). Atendendo ao facto destes algoritmos serem executados em espaços multidimensionais, consideramos as observações x_j e os centroides C_i como vetores definidos, respetivamente, por:

- $x_j = (x_{j1}, x_{j2}, \dots, x_{jp})$;
- $C_i = (c_{i1}, c_{i2}, \dots, c_{ip})$;

Com base nos vetores, o cálculo dos centroides em cada "cluster" é feito da seguinte forma:

$$c_{ik} = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{jk}$$

- c_{ik} representa o valor da k-ésima variável do centroide do i-ésimo "cluster";
- n_i representa o número de observações no i-ésimo "cluster";
- x_{jk} representa o valor da k-ésima variável da j-ésima observação, presente no i-ésimo grupo.

Desta forma, os centroides de cada grupo são representados por um vetor, calculados através dos valores das observações que os constituem (como expresso na fórmula anterior), e o realocar das observações mais distantes, é feito através de uma medida de dissemelhança entre o vetor do centroide e o vetor da observação a realocar. No final são obtidos os k grupos, número previamente definido, constituídos pelas unidades amostrais mais próximas de cada centroide, com base no critério de convergência do método de agrupamento (Ackerman et al., 2010; Jobson, 2012; Kaya Gülağız & Şahin, 2017; Kuo et al., 2002; Sharma, 1995).

Quanto aos métodos de agrupamento hierárquicos, estes algoritmos são referenciados na literatura como uma classe proeminente de técnicas de agrupamento para o estudo comportamental do consumidor. O principal output destes algoritmos consiste num gráfico no formato de árvore, designado dendrograma (Kaya Gülağız & Şahin, 2017; Murtagh & Contreras, 2012). Neste gráfico, o eixo horizontal, composto pelas "folhas" da árvore, tem dispostas as observações (neste caso os sócios), e no eixo vertical estão as distâncias a que os "clusters" foram formados. Os "clusters" são representados pelos nós da árvore, ou seja, cada linha horizontal corresponde a um grupo formado pelos grupos/observações abaixo. O nó superior da árvore é um "cluster" formado por todas as observações, representando a totalidade da amostra. Os restantes "clusters", abaixo

do superior, são tanto mais apurados quanto mais abaixo estiverem na árvore, uma vez que as observações neles contidas são cada vez mais próximas. Tome-se como exemplo o dendrograma relativamente simples, apresentado na Figura 1 com 100 observações, para facilitar a visualização.

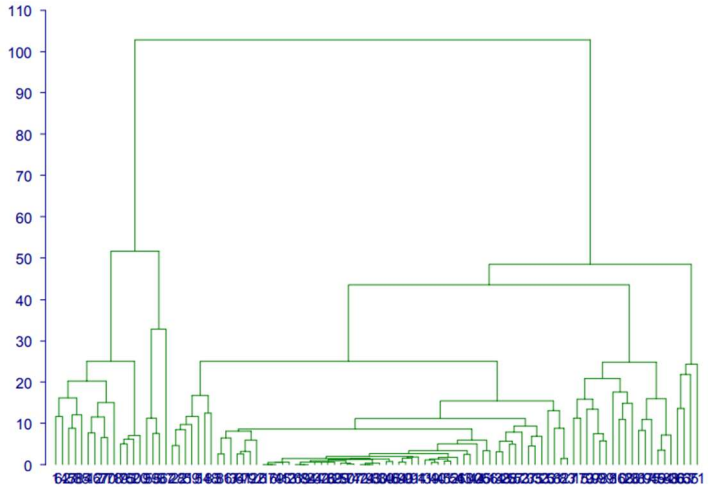

Figura 1 - Dendrograma exemplificativo com 100 observações

No eixo horizontal estão as 100 observações do estudo, e no eixo vertical as distâncias de dissemelhança das quais resultaram os grupos. O utilizador pode optar por definir um critério de máxima distância à qual devem ser formados os grupos, ou, por outro lado, pode realizar uma análise visual ao dendrograma e estabelecer quantos grupos formar, sem utilizar um critério de máxima distância *a priori*. A escolha do número de "clusters" pode ser feita através da observação do dendrograma, tendo presentes os objetivos do estudo e a distância a que são obtidos os "clusters". Também deve ser tida em conta de que forma esta distância tem impacto na heterogeneidade entre os grupos, uma vez que, quanto mais alta a distância, mais baixa a dissemelhança entre os grupos, retirando valor aos resultados. Assim, entende-se que os dendrogramas são outputs que auxiliam bastante a tomada de decisão na análise de grupos. Contudo, estes gráficos são obtidos através de agrupamentos consecutivos, o que gera uma estrutura impraticável em algoritmos não hierárquicos (Schonlau, 2004). Os dendrogramas podem ser construídos através de duas abordagens: por aglomeração e pela abordagem divisiva. A abordagem por aglomeração é designada ascendente, ou de baixo para cima, uma vez que, nesta abordagem, cada observação é apresentada como um "cluster" inicialmente, e a cada iteração do algoritmo, são agrupadas as duas observações mais semelhantes. A abordagem divisiva funciona de maneira oposta. Esta é também designada descendente, ou de cima para baixo, onde todas as observações começam por estar contidas num só "cluster" geral (nó superior da árvore), e são realizadas divisões de forma recursiva, separando um "cluster" em dois e assim sucessivamente, até serem obtidos, novamente, um número de "clusters" igual ao número de observações, dispostas no eixo horizontal (Haque et al., 2022; Kaya Gülağız & Şahin, 2017). Por fim, mencionar que a forma como o dendrograma é construído, isto é, como são definidas quais as

observações a agrupar, depende da técnica de agrupamento designada. Uma vez que neste estudo foram exploradas, principalmente, abordagens dentro dos agrupamentos hierárquicos, decidimos estudar as seguintes técnicas de agrupamento (sendo estas as mais praticadas, segundo a literatura): *single-linkage* (método do vizinho mais próximo); *complete-linkage* (método do vizinho mais afastado); *average-linkage*; e método de Ward (Ackerman & Ben-David, 2016; Jobson, 2012; Sharma, 1995).

A técnica de agrupamento utilizada, para chegar aos resultados finais do projeto, foi o método de Ward. Os restantes métodos convencionais, enumerados previamente, tratam de calcular as distâncias entre os "clusters" e agrupá-los, ou dividi-los, consoante o critério do método utilizado (por exemplo, o *single-linkage* procura a mínima distância possível entre pares de observações). Já o método de Ward, não trata de calcular as distâncias entre as observações, mas sim, o erro da soma dos quadrados (*Error Sum of Squares* (ESS)) provocado pelo possível agrupamento de duas unidades amostrais (ou grupos). Este cálculo é realizado para todas as possibilidades de observações a serem incorporadas nos grupos já existentes, elegendo a observação cuja inclusão, e formação de um novo "cluster", gera o menor aumento possível do erro da soma dos quadrados. O objetivo de minimizar o aumento do ESS está relacionado com a minimização da variância dentro de cada grupo, e não da minimização da variância entre os grupos, como pretendido por outras técnicas. Desta forma, são formados grupos com a maior homogeneidade (intra grupos) possível. O cálculo do erro da soma dos quadrados (ESS) é feito através da fórmula:

$$ESS = \sum_{i=1}^{n_{clus}} \sum_{j=1}^{n_i} \sum_{k=1}^{p} \left(X_{ijk} - \bar{X}_{i \cdot k}\right)^2$$

- n_{clus} representa o número de "clusters" (número inicial igual ao total de observações);
- i representa o índice do "cluster";
- n representa o número de observações para cada "cluster";
- j representa o índice da observação/grupo (dependendo da dimensão) para cada "cluster";
- p representa o número de variáveis (dimensões) na base de dados;
- k representa o índice das variáveis para cada "cluster";
- X_{ijk} representa o valor da *k*-ésima variável da *j*-enésima observação do *i*-enésimo grupo;
- $\bar{X}_{i \cdot k}$ representa a média da *k*-ésima variável, calculada através de todas as observações contidas no *i*-ésimo "cluster".

Assim, o algoritmo reduz o número de grupos, iteração a iteração, agrupando sempre os pares de observações (ou grupos) que minimizem a variância (que é o mesmo que minimizar o aumento do ESS) dentro do novo grupo, criando "clusters" compactos e com observações o mais semelhantes possível (Jobson, 2012; Murtagh & Legendre, 2014; Sharma, 1995).

Para finalizar a Secção das metodologias estudadas, são abordadas as razões pelas quais foram utilizadas determinadas técnicas ao invés de outras, já mencionadas, ao longo deste Capítulo. A combinação destes métodos, ou seja, a utilização do algoritmo hierárquico, recorrendo à matriz de dissemelhança calculada através das distâncias euclidianas, com a aplicação do método de Ward para construir o dendrograma, deu origem aos resultados finais obtidos neste projeto. Outras técnicas também foram estudadas e utilizadas, no entanto, concluiu-se que não seriam apropriadas para alcançar bons resultados. Adicionalmente, os algoritmos não hierárquicos, são os mais utilizados quando se sabe, à partida, o número de grupos desejável. Os algoritmos hierárquicos, ao formar o dendrograma, permitem ao utilizador analisar e entender qual o número de grupos a reter, ou a que distâncias se deve parar de agrupar as observações ("clusters"). O K-Means, é um algoritmo próprio de uma abordagem não hierárquica, o que não se traduz num dos objetivos deste trabalho. Pretende-se sim, desenvolver uma análise e discussão do número definitivo de grupos, com base na inspeção ao dendrograma obtido. Os algoritmos não hierárquicos apresentam uma vantagem quanto aos hierárquicos no que diz respeito à memória utilizada para os executar. Como os algoritmos hierárquicos agrupam as duas observações mais próximas, é necessário testar todos os pares possíveis de ligações para saber qual deles apresenta o valor preferido pela técnica utilizada. Este processo repete-se em cada agrupamento do algoritmo até todas as observações estarem num só "cluster", o que significa um custo de processamento muito alto e a necessidade de grandes quantidades de memória para o executar. Os algoritmos não hierárquicos, por sua vez, priorizam a eficiência computacional, não exigindo uma ferramenta com um desempenho e capacidade de memória tão capaz (Jobson, 2012; Kaya Gülağız & Şahin, 2017; Schonlau, 2004; Sharma, 1995).

Além dos prós e contras já abordados, os algoritmos hierárquicos permitem, através da estrutura do dendrograma, explorar as relações entre as unidades amostrais e grupos, a vários níveis de granularidade (Haque et al., 2022). A granularidade pode traduzir-se na complexidade da interpretação dos dados em diferentes níveis estruturais. Ou seja, num dendrograma executado de forma ascendente (começar com todas as observações e iterar até todas pertencerem ao mesmo grupo), as primeiras camadas junto à base evidenciam uma estrutura granular concreta, uma vez que as observações e grupos já formados apresentam padrões e comportamentos, no contexto deste estudo, muito específicos. Quanto mais alto no dendrograma, menor o número de grupos e, como tal, menos específicos são os padrões captados em cada grupo, ou seja, o dendrograma apresenta, nestas camadas, uma estrutura granular mais abstrata. Na verdade, tanto algoritmos hierárquicos como não hierárquicos são capazes de trabalhar com estruturas de dados de diferentes granularidades, mas só os hierárquicos oferecem uma visualização dessas diferenças da granularidade das estruturas, ao longo do "clustering", através da inspeção ao dendrograma (Feng et al., 2014; Guo et al., 2021; Mccalla et al., 1992). Com base nos objetivos do estudo, nas

características dos dados e nas vantagens e desvantagens das duas classes de algoritmos, foi escolhido o "clustering" hierárquico, para realizar a análise de grupos. Relativamente às técnicas de agrupamento, como referido anteriormente, os métodos estudados, além do método de Ward, não calculam o ESS como este, mas sim as distâncias entre cada par de observações. As distâncias são calculadas da mesma forma, estando as diferenças entre as várias técnicas, no critério definido para escolher o agrupamento a realizar, a cada iteração. Todos os métodos mencionados foram testados, porém, apenas o método de Ward foi considerado viável para prosseguir a uma análise de resultados. As restantes técnicas de agrupamento revelaram uma forte incompatibilidade com a base de dados utilizada, uma vez que ao serem executadas geraram grupos com uma dimensão de apenas uma unidade amostral. As principais razões para estes resultados podem estar relacionadas com a sensibilidade das técnicas a *outliers*, ou, como já mencionado, com a natureza dos dados e propriedades da base de dados deste estudo. Com estes resultados a análise de grupos era impraticável, pelo que a técnica de agrupamento designada foi o método de Ward, gerando resultados muito mais favoráveis e admissíveis para prosseguir a análise.

3.3. Power BI e visualização de dados

No desenvolvimento de variáveis, como já referido, foi utilizado o Power BI, nomeadamente a linguagem de programação DAX. As variáveis definidas foram desenvolvidas através desta linguagem, adicionando novas colunas à tabela "Conta" no Power BI Desktop, com o auxílio de ferramentas como funções, filtros, condições e variáveis auxiliares. Também foram aplicados alguns filtros na Power Query, de modo a facilitar o trabalho no Power BI Desktop e minimizar o peso do ficheiro.

Já com todas as variáveis do modelo implementadas, foi necessário aplicar as restrições e definir a amostra dos 20'000 sócios. Foi gerada uma nova tabela no Power BI, através da criação de um objeto fictício com o nome "*Sample*" (amostra). Este é um objeto que, no painel de "Vista do modelo", encontra-se sem qualquer ligação aos restantes objetos, sendo apenas utilizado para gerar a amostra e incluir somente as 20'000 observações aleatórias, mediante as restrições estabelecidas. A criação desta amostra consistiu em adicionar uma nova tabela no Power BI e, através de linguagem DAX, foram utilizadas duas funções: *Calcultable* e *Sample*. A função *Sample* gera uma amostra aleatória de um dado número de observações de uma determinada tabela, já presente no ficheiro do Power BI. Assim, nos seus argumentos, foram definidos o número "20000", que representa a dimensão da amostra, e a tabela "Conta", onde estão todas as observações da população de sócios. A função *Calculatable* gera uma tabela modificada por filtros. Neste caso, nos seus argumentos, a tabela será "Sample" e os filtros serão as condições impostas pelas três variáveis, já presentes no modelo, que correspondem às restrições da amostra.

Assim, foi possível obter uma tabela com 20’000 observações, que respeitam as três restrições definidas. É através do objeto “Sample” que é construída a tabela, em Power BI Desktop, com todas as variáveis (colunas) que desejemos incluir no modelo, sendo esta a base de dados final exportada do Power BI e utilizada no R Studio para testar os algoritmos de segmentação.

Após executar os algoritmos de segmentação, e tendo chegado à abordagem definitiva, foram gerados os cinco grupos no R Studio, sob a forma de tabelas idênticas à base de dados final. Estas tabelas apresentam a mesma informação que a base de dados carregada para o R Studio, com a diferença de que contêm apenas as observações do grupo que representam. Assim, através da separação dos grupos nas cinco tabelas, foi possível extraí-las, separadamente, para o Power BI, novamente. No Power BI, o objetivo nesta fase seria estudar os grupos de forma isolada com recurso a técnicas de visualização, como já referido, através da elaboração de *dashboards*. Estes devem apresentar um determinado conjunto de variáveis (enumeradas e exploradas no Capítulo seguinte), que abordem, de forma geral, as várias unidades de negócio que o modelo pretende estudar.

Para desenvolver estes gráficos, foi necessária a elaboração de variáveis auxiliares de dois tipos: variáveis que definam intervalos de valores, e variáveis que estabeleçam uma ordem para estes intervalos. As primeiras servem para conseguirmos incluir as observações dentro dos intervalos que definimos, caso contrário os gráficos apresentariam cada observação individualmente, dificultando a análise visual pela informação disposta numa maneira mais densa e saturada. Vejamos o exemplo da variável “Valor total de encomendas realizadas”, para a tabela da amostra. Na Figura 2 está representada esta variável, contando com uma observação para cada sócio, vistas individualmente, dificultando a perceção de como esta variável está distribuída pela amostra.

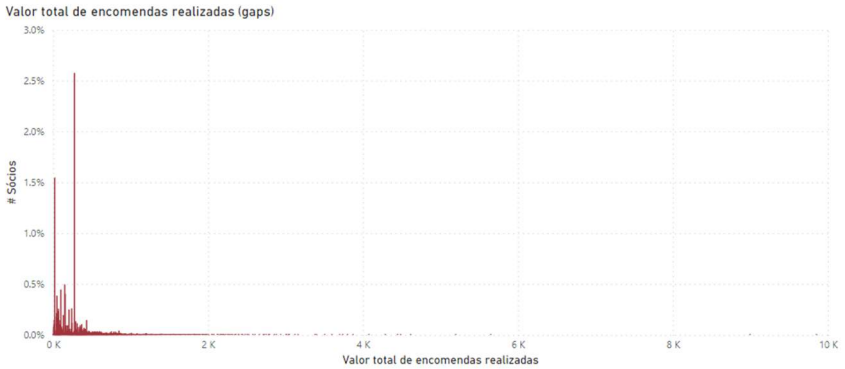

Figura 2 - Variável "Valor total de encomendas realizas" para a amostra, sem intervalos de valores

Relativamente às variáveis auxiliares para estabelecer a ordem correta dos intervalos, estas são necessárias, uma vez que, por defeito, o Power BI organiza os intervalos através da ordem dos algarismos, individualmente. Na Figura 3, está representada a ordenação ascendente feita pelo Power BI, sem a variável auxiliar, onde está representada esta ordenação, algarismo a algarismo.

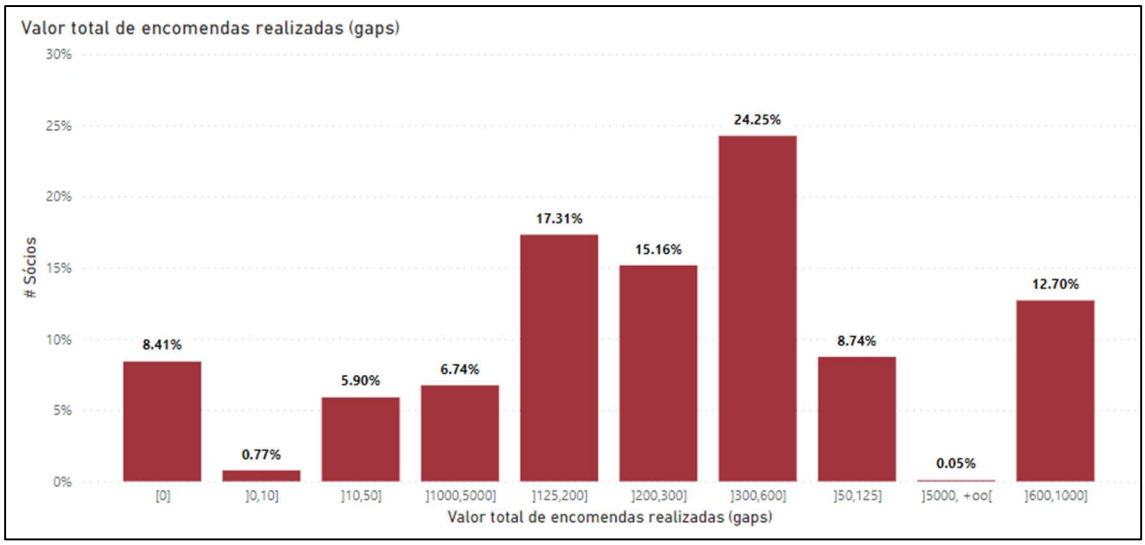

Figura 3 - Variável "Valor total de encomendas realizas" para a amostra, com ordenação incorreta dos intervalos

Assim, foi necessário desenvolver variáveis auxiliares para cada variável com representação gráfica, e para cada um dos *dashboards* (no total seis *dashboards*, um para a amostra e cinco para os grupos, individualmente).

Na Figura 4, as observações estão repartidas pelos intervalos definidos, com a ordenação correta, contribuindo para uma análise visual mais capaz, à distribuição da variável.

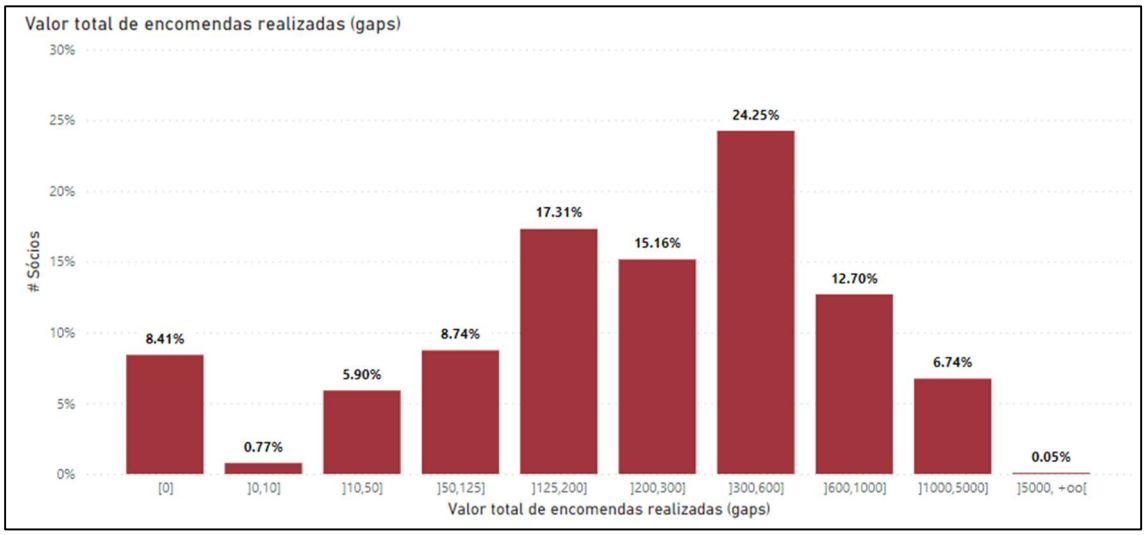

Figura 4 - Variável "Valor total de encomendas realizas" para a amostra, com intervalos de valores e ordenação correta

4. Análise e discussão dos resultados

Neste Capítulo é apresentada a análise e a discussão dos resultados do projeto, começando por uma apreciação geral da amostra, seguida por uma análise mais detalhada a cada grupo e, finalmente, comparações entre eles, de modo a vincar as características que os diferenciam.

4.1. Análise da amostra e apresentação de indicadores

Iniciando pela amostra, como já referido, esta contempla 20’000 sócios, todos cumpridores das três restrições mencionadas no Capítulo 2, e respeitante a consumos ao longo das épocas desportivas 2021/2022 e 2022/2023. Para analisar a amostra e os grupos obtidos, foram utilizadas duas ferramentas para visualizar os dados: as diferenças entre as médias dos grupos para cada variável, feito em R, e *dashboards* no Power BI, que desenvolvemos especificamente para este estudo, com os principais indicadores. Quanto aos *dashboards* em Power BI, foi selecionado um conjunto de indicadores (variáveis-chave), considerados os mais pertinentes para estudar o comportamento dos sócios pertencentes ao grupo (ou neste caso, à amostra). Estes indicadores foram:

- País;
- Quotas em dia;
- Género;
- Registado no site;
- Faixa etária;
- Antiguidade de sócio;
- Sócio já adquiriu produtos de bilhética;
- Sócios já adquiriu produtos de *merchandising*;
- Detentor Red Pass futebol;
- Aderente “Mais Vantagens”;
- Aderente “Solução Família”;
- Valor total de encomendas realizadas;
- Valor pago em quotas;
- Valor de quotas em dívida;
- Valor total gasto em *merchandising*;
- Assiduidade ao longo da época (jogo a jogo);

- Assiduidade Red Pass;
- Valor total gasto em Red Pass;
- Somatório do valor dos bilhetes adquiridos para futebol;
- Saldo "Mais Vantagens" acumulado;
- Longevidade da adesão às "Mais Vantagens", em meses;
- Número de vezes que sócio visitou o site;
- Longevidade do registo do sócio no site.

Apesar do modelo ser composto por muitas outras variáveis, este grupo de indicadores abrange todas as categorias (recorde-se que as categorias são aquelas definidas no Capítulo 2, referentes às unidades do negócio) que se pretendem estudar, facilitando a análise aos sócios de cada grupo, sem que seja necessário estudar todas as variáveis. Assim, é possível perceber de forma mais imediata quais os grupos compostos por sócios mais ligados ao clube, e aqueles constituídos por sócios com uma pior relação com o Sport Lisboa e Benfica.

Na Figura 5, estão representados alguns dos elementos visuais que desenvolvemos através do Power BI, com as variáveis mencionadas, para os 20'000 sócios da amostra.

Figura 5 - Indicadores gerais para análise da amostra

Este *dashboard* remete para estatísticas mais gerais que procuram analisar dados demográficos, como a faixa etária ou o género; relação entre os sócios e as diferentes áreas do clube, como a aquisição de "Red Pass" ou a adesão ao programa "Mais Vantagens"; e a existência, ou não, de consumos nos dois principais ramos de vendas, a aquisição de *merchandising* ou de bilhética.

Através da visualização destes gráficos podemos concluir que a grande maioria dos sócios presentes nesta amostra são homens portugueses, predominantemente com idades entre os 24 e 54 anos. Quanto às várias áreas do negócio, por um lado, a maior parte tem os pagamentos de quotas em dia, está registada no site e já adquiriu produtos de bilhética. Por outro lado, verifica-se uma menor relação dos sócios com outras áreas de consumo, por exemplo a baixa aquisição de produtos de *merchandising*, ou a fraca adesão ao "Red Pass", plano "Mais Vantagens", ou programa "Solução Família". Por fim, a antiguidade de sócio é um dado pouco conclusivo, uma vez que não apresenta uma tendência ou um valor bastante predominante em relação aos restantes.

Além destes indicadores, também são exploradas outras variáveis com o objetivo de estudar os consumos dos sócios em vários setores do clube. Na Figura 6 estão representadas essas variáveis.

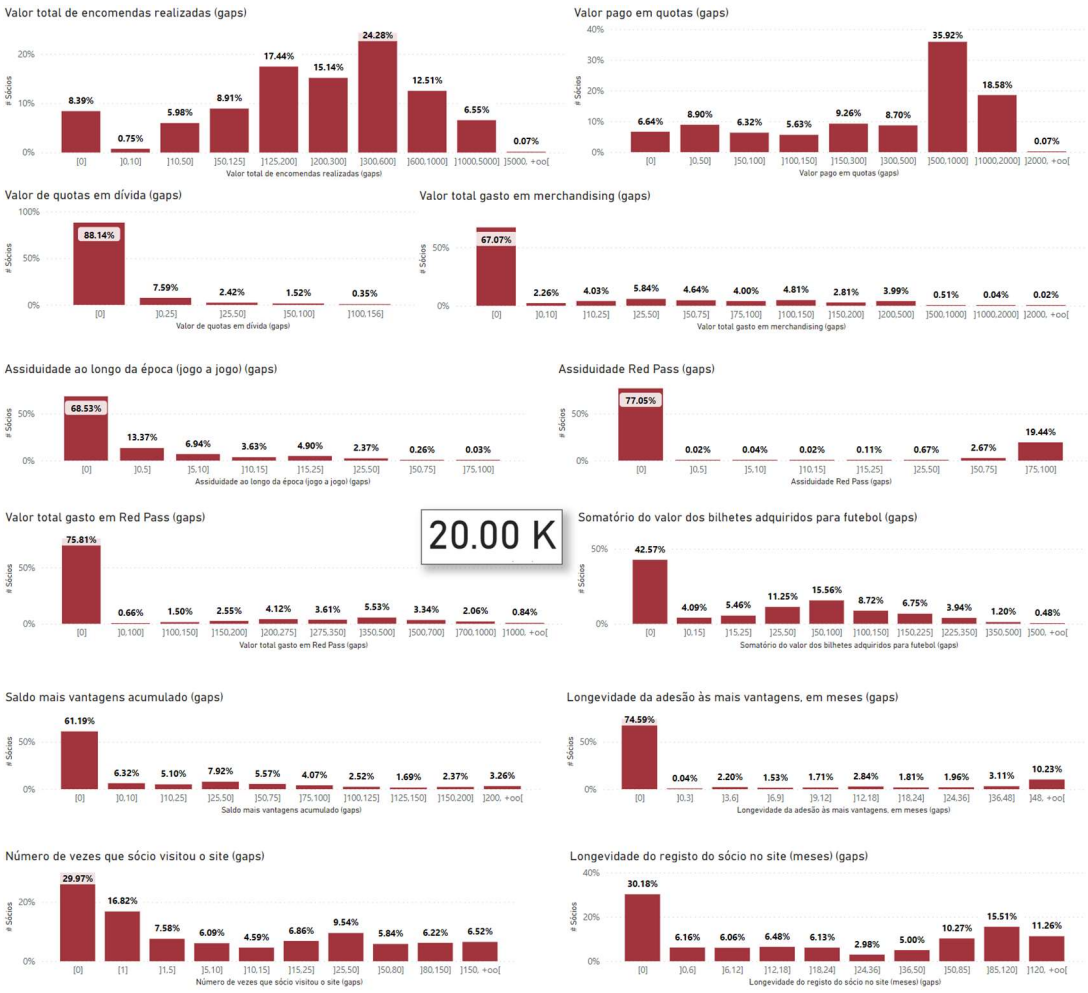

Figura 6 - Indicadores referentes a consumos para análise da amostra

Através destes dados, é possível perceber que, na maioria das variáveis, a maioria dos sócios não apresentam qualquer valor de consumo nas várias áreas do negócio, o que vai de encontro à informação retirada dos primeiros gráficos. Ainda assim, é possível desenvolver um pouco estas conclusões em relação a algumas variáveis. Quanto ao valor total das encomendas realizadas, verifica-se que mais de 50% dos sócios gastou entre 125€ e 600€, ao longo do horizonte temporal estipulado. No entanto, este valor é fortemente impactado pelos pagamentos de quotas, como é possível verificar no gráfico seguinte, onde uma percentagem significativa de sócios realizou pagamentos entre 500€ e 2,000€. O gráfico que ilustra a assiduidade através do uso de "Red Pass" apresenta uma percentagem elevada de sócios com 0% de assiduidade, representativa dos sócios que não detêm "Red Pass", mas também uma parcela considerável de sócios com assiduidades entre 75% e 100%, face aos restantes intervalos. Quantos às variáveis que respeitam à utilização do site oficial do Sport Lisboa e Benfica ("Número de vezes que sócios visitou o site" e "Longevidade do registo do sócio no site (meses)"), pode verificar-se que, apesar de existir um número significativo de sócios sem qualquer interação, são variáveis com os valores bem distribuídos entre os intervalos, não existindo um ou mais intervalos predominantes.

Assim, pode concluir-se que, na globalidade da amostra, existem muitos mais sócios ausentes do clube, do que sócios verdadeiramente ligados ao mesmo. Desta forma, espera-se que os grupos de sócios pouco relacionados tenham uma dimensão muito maior do que os grupos constituídos por sócios com melhores relações, com o Sport Lisboa e Benfica.

Após a criação dos grupos, através dos algoritmos de "clustering" hierárquico do R Studio, e da tabela de comparação das médias, foram isoladas sete variáveis, consideradas pertinentes do ponto de vista do negócio, para uma análise geral dos grupos. Na Tabela 2, estão apresentadas as médias referentes a estas variáveis para os cinco grupos, com os valores dispostos numa escala de cores, onde, no limite, o vermelho representa um comportamento típico dos sócios com pior relação com o clube, e os valores a verde simbolizam o oposto. A escala de cores é calculada com base nos valores máximos e mínimos de cada variável, ou seja, com base nos valores possíveis dentro da amplitude de cada uma.

Tabela 2 - Tabela de médias das variáveis gerais para os 5 grupos

	C1	C2	C3	C4	C5
Antiguidade de sócio (anos)	3.44	13.41	20.39	17.49	21.77
Última entrada no estádio	4.46	2.94	1.20	11.13	1.72
Número de encomendas realizadas	5.42	19.38	19.86	10.84	24.32
Número de transações em merchandising	0.41	1.63	2.03	0.46	3.96
Número total de bilhetes adquiridos	4.75	33.02	50.08	4.72	74.42
Valor total gasto em Red Pass	17.24	239.60	649.45	18.53	1315.97
Percentagem de sócios com quotas em dia	85.51%	97.30%	99.31%	95.32%	96.26%
Dimensão (# sócios)	7952	3445	1015	7374	214

Através da escala de cores, entende-se que os grupos de sócios mais desligados do clube são os grupos 1 e 4 (designados C1 e C4), enquanto os grupos 3 e 5 (designados C3 e C5) são compostos por sócios que apresentam uma relação mais presente com o Sport Lisboa e Benfica. De realçar também o caso do grupo 2 (representado por C2), que não fará sentido incluir em nenhum destes dois leques, uma vez que contém valores intermédios para as variáveis apresentadas.
A Tabela 3 representa, sucintamente, a relação de cada grupo com o clube, assim como a sua dimensão, ou seja, o número de sócios que constituem o grupo.

Tabela 3 - Relação dos grupos com o clube e respetiva dimensão

	C1	C2	C3	C4	C5
Relação com o clube	Fraca	Moderada	Forte	Fraca	Forte
Dimensão (# sócios)	7952	3445	1015	7374	214
Dimensão (%)	39.76%	17.23%	5.08%	36.87%	1.07%

Através da Tabela 3, e em conformidade com os valores de consumo analisados, do total da amostra, é possível verificar que os grupos de sócios com uma fraca relação com o clube constituem mais de 75% das observações (15,326 sócios). Os grupos onde prevalece uma forte relação, entre os sócios e o clube, são constituídos por 1,229 sócios, e os restantes 3,445 fazem parte do grupo 2, onde a relação dos sócios com o clube é classificada como moderada.

4.2. Análise e comparação de grupos

Nesta Secção do Capítulo 4 serão analisados, individualmente os cinco grupos, através da seguinte estrutura: primeiro são analisados os *dashboards*, para cada grupo, obtidos através do Power BI, à semelhança de como foi feita a análise da amostra, e evidenciados os valores que prevalecem em relação aos outros grupos; de seguida, são observados os valores médios para todas as variáveis, e desta forma, são escolhidos indicadores específicos para cada grupo, de modo a destacar as suas características em relação aos restantes. Finalmente, e com base nas duas etapas anteriores, são definidas as características e comportamentos mais vincados, pelos sócios de cada grupo.

Ainda neste Capítulo, e posteriormente à análise dos grupos de forma individual, serão realizadas comparações mais minuciosas entre os grupos 1 e 4 e também entre os grupos 3 e 5, uma vez que, estes pares de grupos têm algumas semelhanças entre si. Assim, é possível entender as principais particularidades entre os sócios que têm pior relação com o clube (grupos 1 e 4), e os sócios mais ligados (grupos 3 e 5). O grupo 2 não fará parte desta análise, uma vez que é o único grupo cuja relação dos sócios com o clube é considerada "moderada".

4.2.1. Grupo 1

Começando pelo grupo 1, foram analisados os gráficos desenvolvidos através do Power BI, respeitantes a dados demográficos e à visão geral do consumo, presentes na Figura 7.

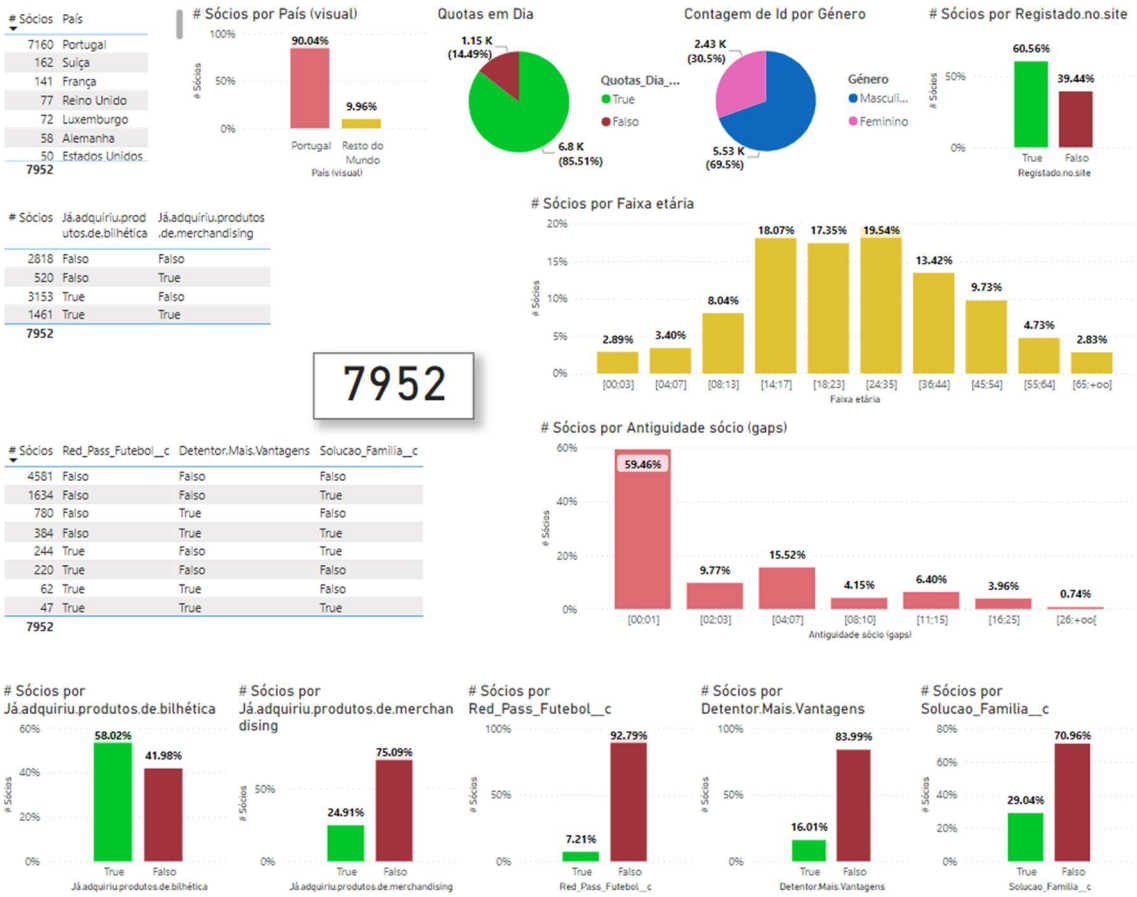

Figura 7 - Indicadores gerais para análise do grupo 1

Através destes gráficos, e da comparação com os restantes grupos, é possível concluir que este é o grupo com maior percentagem de sócios:

- Fora de Portugal;
- Do sexo feminino;
- Com quotas em atraso;
- Não inscritos no site;
- Com uma antiguidade até 12 meses;
- Que não adquiriram produtos de *merchandising* no horizonte temporal;
- Não aderentes ao programa "Mais Vantagens".

Passando agora para os dados referentes a consumos nas várias áreas do negócio, obtiveram-se os gráficos presentes na Figura 8.

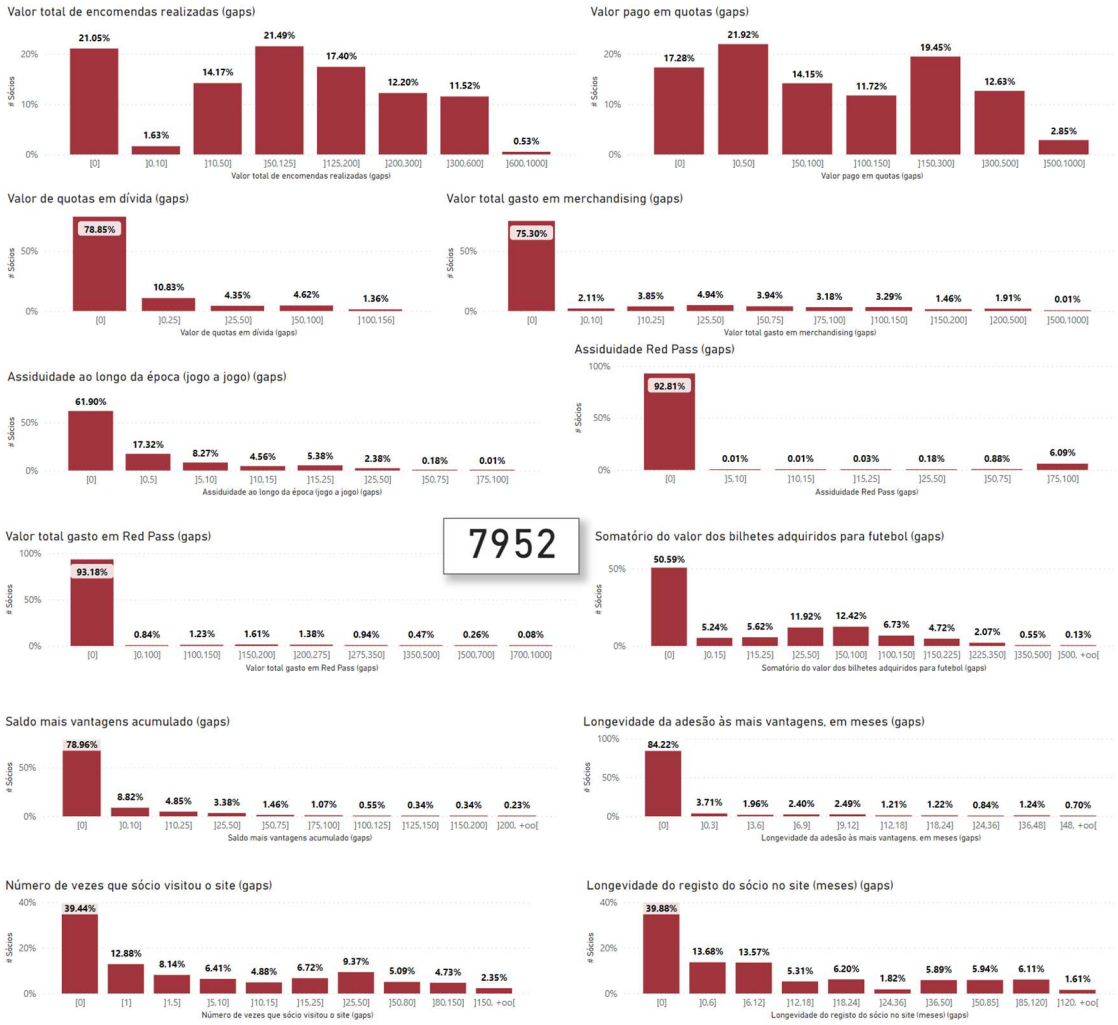

Figura 8 - Indicadores referentes a consumos para análise do grupo 1

Com base nestes gráficos, e em comparações com os gráficos dos outros grupos, é possível concluir que este é o grupo com maior percentagem de sócios que:

- Não registaram qualquer valor de consumo através de encomendas;
- Não registaram qualquer valor de consumo através de *merchandising*;
- Nunca acumularam saldo em carteira virtual.

Fazendo uma análise mais exaustiva através das comparações entre médias, optou-se por selecionar as variáveis apresentadas na Tabela 4 para auxiliar a caracterização dos sócios deste grupo.

Tabela 4 - Tabela de médias das variáveis selecionadas para a definição do grupo 1

	C1	C2	C3	C4	C5
Idade	27.99	40.18	50.29	47.41	48.72
Antiguidade de sócio (anos)	3.44	13.41	20.39	17.49	21.77
Diferencial entre número na presente época e número médio por época (encomendas)	0.91	0.81	-0.24	0.03	0.46
Valor de quotas pago	132.32	694.97	875.27	784.87	864.99
Número de encomendas (quotas)	3.92	10.81	11.81	9.50	8.20
Frequência de pagamentos, na presente época (quotas)	2.49	5.53	5.75	4.66	3.91
Número de transações nas lojas físicas	3.72	12.64	15.97	9.83	12.31
Presenças ao longo da época (jogo a jogo)	0.96	1.03	0.06	0.77	0.62
Antiguidade de Red Pass	0.27	5.17	9.77	1.34	9.79
Número de redenções	0.15	1.52	2.19	0.31	3.25
Valor total de redenções (saldo utilizado)	2.48	36.91	83.39	7.90	166.24
Saldo acumulado em carteira virtual	6.60	67.04	144.23	18.48	230.62
Longevidade do registo do sócio no site (meses)	18.81	65.15	61.53	51.51	60.95

Através destas variáveis, foram retiradas as seguintes conclusões:

- A idade e antiguidade de sócio neste grupo são muito inferiores face à generalidade. No caso concreto da antiguidade, mais de 50% dos sócios deste grupo, ainda não atingiram a marca dos 12 meses;
- O diferencial entre o número de encomendas na presente época e o número médio de encomendas por época é o maior de todos os grupos. Este dado indica que a relação dos sócios com o clube melhorou da época 2021/2022 para a época 2022/2023. A justificação mais aceite é a de que a maior parte destes sócios só realizaram a sua adesão na época mais recente, não apresentando consumos na época anterior;
- Outros fatores que evidenciam o baixo valor da antiguidade dos sócios são os indicadores respeitantes ao pagamento de quotas. Estes são os sócios com os valores, número e frequência de pagamentos mais baixos;
- Este grupo também remete para números significativamente baixos nas transações realizadas em lojas físicas;
- Relativamente à assistência aos jogos de futebol, este grupo apresenta dos valores mais altos comparativamente aos restantes, porém, uma presença por sócio, em média, ao longo de uma época, representa um valor consideravelmente baixo (os sócios mais relacionados com o clube têm este valor próximo de zero porque a grande maioria assiste aos jogos através da utilização de "Red Pass", e não da aquisição de bilhetes "jogo a jogo");

- Quanto ao Red Pass, este grupo apresenta valores também muito baixos, destacando o histórico de Red Pass, ou seja, o número de épocas em que os sócios garantiram a sua aquisição, que é muito inferior a todos os outros;
- Observando as 3 variáveis respeitantes ao programa "Mais Vantagens", depreende-se que estes são os sócios menos envolvidos com esta área do negócio;
- Por fim, a utilização do site apresenta valores semelhantes aos outros grupos, com exceção da longevidade do registo, que é muito baixa. Novamente, tanto a baixa antiguidade de "Red Pass" como a do registo no site, são fortemente impactadas pela baixa antiguidade dos sócios.

Através destes dados, as conclusões que podemos retirar face ao comportamento dos sócios deste grupo são:

→ Em termos demográficos é o grupo que inclui uma maior diversidade de sócios, nomeadamente com o país e o género;

→ É o grupo com mais sócios com quotas em atraso;

→ São os sócios menos ligados ao *merchandising* e ao programa "Mais Vantagens";

→ Sócios muito recentes, cuja maioria não aderiu há mais de 12 meses.

4.2.2. Grupo 4

Após a análise ao grupo 1 será agora estudado o grupo 4, para posteriormente realizar as comparações entre eles. Para tal, recorre-se novamente aos gráficos do Power BI. Na Figura 9 são apresentadas as estatísticas demográficas e os dados gerais de consumo, referentes ao grupo 4.

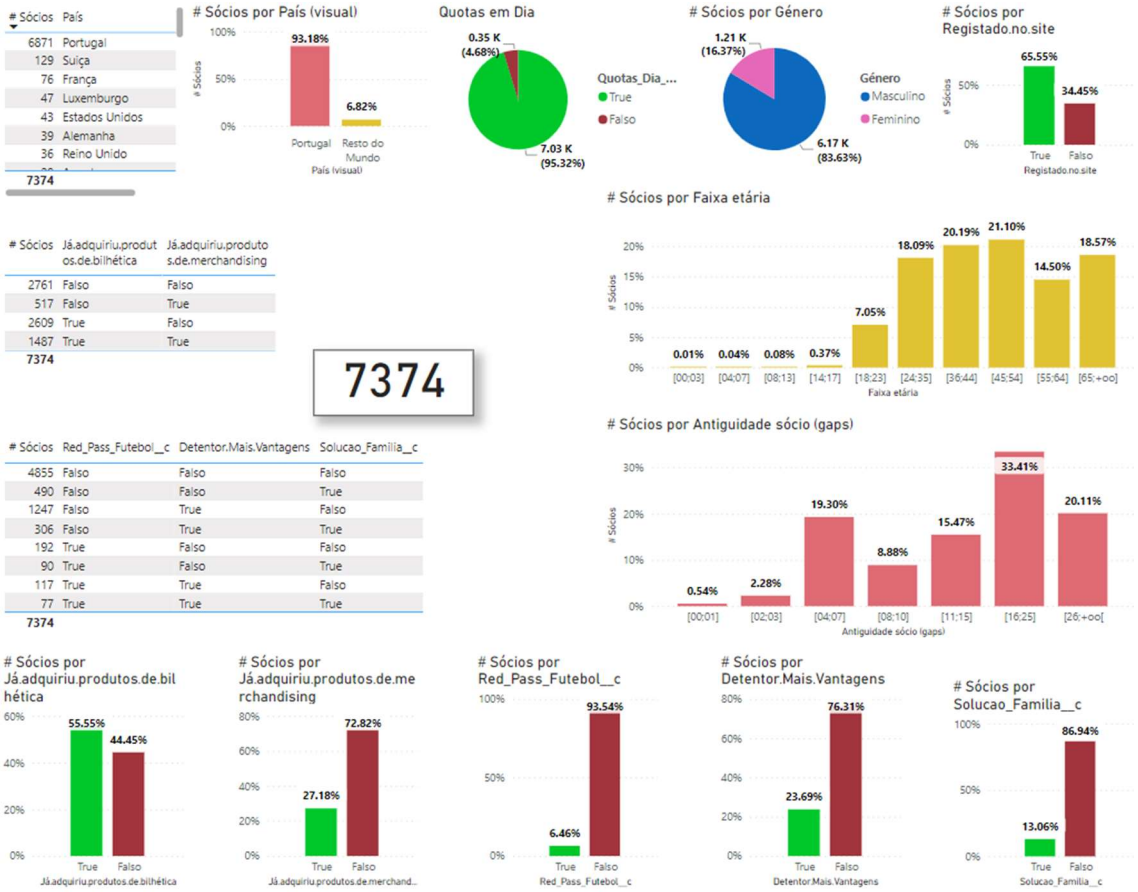

Figura 9 - Indicadores gerais para análise do grupo 4

Através destes gráficos, e da comparação com os restantes grupos, é possível concluir que este é o grupo com maior percentagem de sócios:

- Com antiguidade entre os 16 e os 25 anos;
- Sem artigos de bilhética adquiridos no horizonte temporal;
- Sem "Red Pass";
- Sem adesão à "Solução Família".

Para estudar com maior detalhe os diversos consumos, foram desenvolvidos os gráficos ilustrados na Figura 10.

Figura 10 - Indicadores referentes a consumos para análise do grupo 4

Através da observação destes dados, e comparação com os outros grupos, foi possível inferir que este é o grupo com maior percentagem de sócios:

- Com consumos gerais (encomendas) cujos valores estão compreendidos entre 125€ e 600€;
- Com um valor de quotas pago entre 500€ e 1000€;
- Com 0% de assiduidade no "Red Pass";
- Com uma única visita ao site (possivelmente a visita onde fizeram a inscrição).

Para analisar mais pormenorizadamente o comportamento dos sócios deste grupo, foram selecionadas as variáveis pertencentes à Tabela 5, juntamente com as suas médias, para todos os grupos.

Tabela 5 - Tabela de médias das variáveis selecionadas para a definição do grupo 4

	C1	C2	C3	C4	C5
Antiguidade de sócio (anos)	3.44	13.41	20.39	17.49	21.77
Última entrada no estádio	4.46	2.94	1.20	11.13	1.72
Número de encomendas realizadas na presente época	3.62	10.50	9.69	5.45	12.62
Valor de quotas pago	132.32	694.97	875.27	784.87	864.99
Número de membros "Solução Família"	1.64	1.51	1.71	1.27	1.77
Longevidade da adesão à "Solução Família", em meses	7.15	7.07	11.38	4.03	7.79
Número de transações em merchandising	0.41	1.63	2.03	0.46	3.96
Número total de bilhetes adquiridos	4.75	33.02	50.08	4.72	74.42
Número de jogos assistidos com Red Pass	1.51	14.00	21.84	1.31	20.18
Valor gasto em Red Pass, na presente época	12.83	147.32	391.48	11.03	793.43
Número de bilhetes adquiridos para modalidades	0.45	2.07	3.37	0.39	9.44
Número de modalidades assistidas	0.74	1.54	1.54	0.66	1.85
Número total de bilhetes para visitas ao estádio/museu	0.10	0.23	0.16	0.08	0.31
Número de redenções	0.15	1.52	2.19	0.31	3.25
Saldo acumulado	6.60	67.04	144.23	18.48	230.62
Valor de desconto direto obtido em parceiros	5.59	25.63	33.68	26.95	39.40
Longevidade do registo do sócio no site (meses)	18.81	65.15	61.53	51.51	60.95

Através destes dados concluiu-se que:

- A antiguidade de sócio destaca-se como um valor elevado e próximo dos grupos de sócios mais ativos;
- Por outro lado, este grupo apresenta o maior valor na variável "Última entrada no estádio", indicativo de que estes sócios são, provavelmente, aqueles que menos frequentam o estádio. Este valor torna-se ainda mais significativo comparando a discrepância para com os restantes grupos;
- Pelo baixo número de encomendas, entende-se que são sócios sem o hábito de realizar consumos nas várias áreas do negócio, sendo o pagamento de quotas, possivelmente, o consumo predominante neste grupo;
- O valor de quotas pago é o terceiro maior, o que é justificado pela antiguidade de sócio, mas também pelo ponto anterior;
- As variáveis da "Solução Família" demonstram que este é o grupo menos ligado a esta área do negócio;
- Quanto ao número de encomendas do tipo *merchandising*, apresenta valores muito inferiores face aos grupos de sócios mais relacionados;
- As variáveis de maior destaque neste grupo correspondem aos indicadores das áreas de bilhética. Pelo número de bilhetes adquiridos, assistências a jogos com "Red Pass", incluindo valor gasto no mesmo, número de modalidades assistidas, e visitas ao estádio ou museu, entende-se que este é o grupo com os consumos de bilhética mais baixos;

- Quanto às variáveis referentes ao plano "Mais Vantagens", este grupo também apresenta valores significativamente baixos, exceto na variável "Valor de desconto direto obtido em parceiros". Isto deve-se ao facto de muitos destes sócios frequentarem a Repsol (um dos poucos parceiros do Sport Lisboa e Benfica que oferece desconto direto além do saldo acumulado em carteira virtual), apesar de não demonstrarem interesse em usufruir das vantagens deste programa noutros parceiros;
- Por fim, relativamente ao site, o número de visitas é baixo, como já referido, porém, a longevidade do registo no site apresenta um valor muito elevado, à semelhança dos "clusters" mais envolvidos. Este facto pode ser explicado pela antiguidade dos sócios deste grupo que, não necessitam de frequentar regularmente o site para apresentar uma data de inscrição mais antiga.

Através de uma avaliação a todos estes dados, pode concluir-se que as características que melhor definem os sócios do grupo 4 são:

→ São sócios com uma fraca ligação ao clube, em geral, apesar de apresenteram uma antiguidade considerável quanto à sua adesão;

→ São aqueles que menos frequentam o estádio, e consequentemente, os menos presentes nos jogos, nas várias modalidades;

→ Os consumos são constituídos, principalmente, por pagamentos de quotas.

4.2.3. Principais diferenças entre os grupos 1 e 4

Feita a análise individual a estes dois grupos, o objetivo seguinte é perceber as principais diferenças entre eles. Para tal, foram selecionadas as variáveis consideradas mais determinantes para esta análise, e construída a Tabela 6.

Tabela 6 - Tabela de médias das variáveis selecionadas para a comparação entre os grupos 1 e 4

	C1	C2	C3	C4	C5
Idade	27.99	40.18	50.29	47.41	48.72
Antiguidade de sócio (anos)	3.44	13.41	20.39	17.49	21.77
Última entrada no estádio	4.46	2.94	1.20	11.13	1.72
Valor total de encomendas realizadas	129.59	745.91	1185.07	291.12	2637.78
Longevidade da adesão à solução família, em meses	7.15	7.07	11.38	4.03	7.79
Valor total gasto em merchandising	19.54	84.99	110.94	20.37	257.12
Número de transações no canal site	1.11	4.36	2.53	0.73	9.12
Número de transações no canal APP	0.47	1.49	0.91	0.21	1.54
Número de transações no canal loja física	3.72	12.64	15.97	9.83	12.31
Presenças ao longo da época (jogo a jogo)	0.96	1.03	0.06	0.77	0.62
Número de modalidades assistidas	0.74	1.54	1.54	0.66	1.85
Saldo acumulado	6.60	67.04	144.23	18.48	230.62
Longevidade da adesão às Mais Vantagens, em meses	2.23	15.16	16.86	8.92	13.93
Longevidade do registo do sócio no site (meses)	18.81	65.15	61.53	51.51	60.95

Através da análise a estes indicadores, é possível tirar as seguintes conclusões:

- → A variável "Idade" é uma das que mais se destaca pela discrepância entre os valores médios dos dois grupos. Verifica-se que, em média, os sócios do grupo 4 apresentam um valor, para esta variável, muito mais alto que os sócios do grupo 1;
- → Esta diferença da "Idade" também contribui para valores muito distantes em relação à variável "Antiguidade de sócio". O grupo 4 tem um valor médio bastante elevado, ao contrário do grupo 1 que, como já mencionado, é composto por uma maioria de sócios com uma antiguidade igual ou inferior a 12 meses;
- → O indicador da antiguidade de sócio explica a disparidade entre estes dois grupos, nas variáveis referentes a longevidade de adesão/inscrição, para as várias áreas do negócio. Observem-se as variáveis "Longevidade da adesão à solução família, em meses" e "Longevidade da adesão às Mais Vantagens, em meses";
- → Quanto a matéria de consumos, através das variáveis "Valor total de encomendas realizadas" e "Saldo acumulado", são demonstrados valores mais altos para o grupo 4, embora, no que respeita a consumos, especificamente, de *merchandising*, sejam apresentados valores muito semelhantes para os dois grupos;
- → Através dos canais de venda, depreende-se que os sócios do grupo 1 têm muito mais aptidão para utilizar meios digitais do que os sócios do grupo 4, que preferem dirigir-se às lojas físicas (dado que pode estar diretamente relacionado com as idades médias dos dois grupos);

→ Por fim, com base nas variáveis "Última entrada no estádio", "Presenças ao longo da época (jogo a jogo)" e "Número de modalidades assistidas", pode concluir-se que os sócios do grupo 1 frequentam muito mais o estádio que os sócios do 4.

4.2.4. Grupo 2

Passando para o grupo 2, o único onde a relação entre os sócios e o clube é considerada "moderada", a análise foi, novamente, iniciada pelos *dashboards* desenvolvidos no Power BI. Os primeiros elementos visuais, presentes na Figura 11, correspondem aos dados demográficos e gerais de consumo.

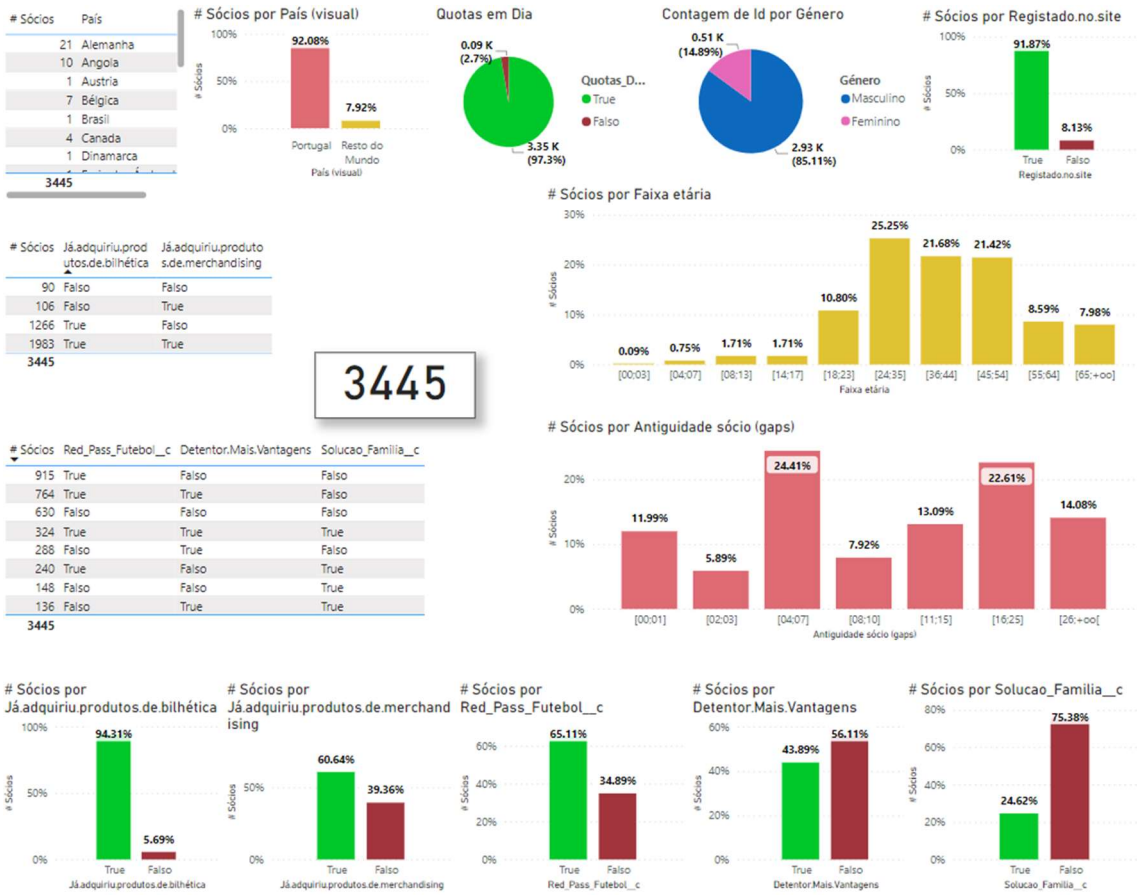

Figura 11 - Indicadores gerais para análise do grupo 2

Através da observação destes dados, e comparando-os com os restantes grupos, é possível inferir que este é o grupo com maior percentagem de sócios:

- Registados no site;
- Com idades entre 24 e 35 anos.

Para observar mais detalhadamente os consumos nas diversas áreas do negócio, elaborámos os gráficos presentes na Figura 12.

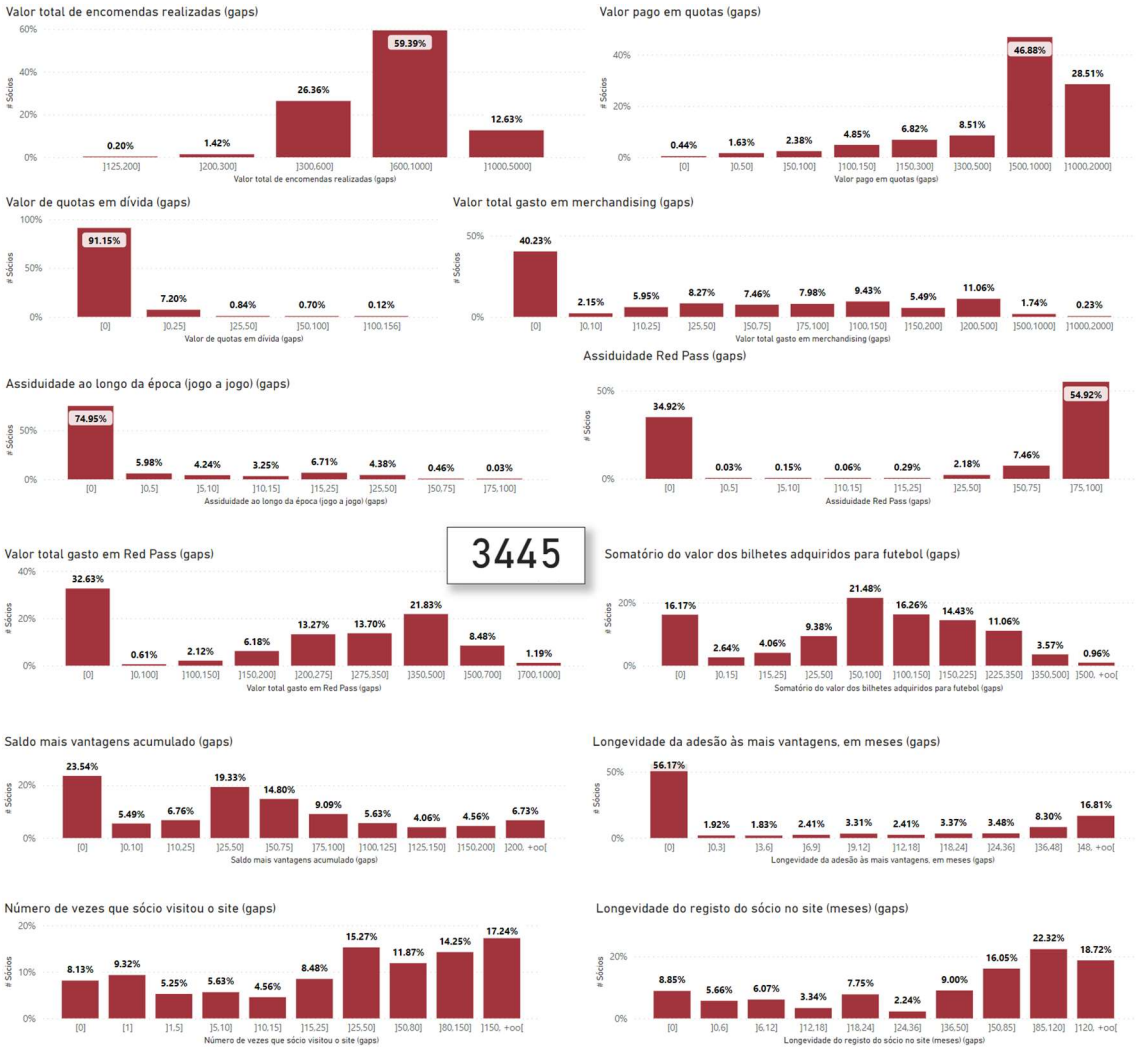

Figura 12 - Indicadores referentes a consumos para análise do grupo 2

Com recurso à visualização destes dados, concluiu-se que este é o grupo com maior percentagem de sócios:

- Com o valor das encomendas (consumos totais) entre 600€ e 1,000€;
- Com consumos de "Red Pass" entre 350€ e 500€;
- Com saldo acumulado em carteira virtual, com valores entre 25€ e 50€;
- Inscritos no site há mais de 85 meses (7 anos).

Finalmente, foram selecionadas variáveis em concreto, da tabela de médias, consideradas fundamentais para auxiliar a caracterização dos sócios deste grupo. As variáveis em destaque estão contidas na Tabela 7.

Tabela 7 - Tabela de médias das variáveis selecionadas para a definição do grupo 2

	C1	C2	C3	C4	C5
Idade	27.99	40.18	50.29	47.41	48.72
Última entrada no estádio	4.46	2.94	1.20	11.13	1.72
Número de encomendas realizadas	5.42	19.38	19.86	10.84	24.32
Valor total de encomendas realizadas	129.59	745.91	1185.07	291.12	2637.78
Diferencial entre valor na presente época e valor médio por época (encomendas)	24.17	36.65	-52.63	4.06	-90.77
Número de transações em merchandising	0.41	1.63	2.03	0.46	3.96
Número de transações no canal site	1.11	4.36	2.53	0.73	9.12
Número de transações no canal APP	0.47	1.49	0.91	0.21	1.54
Número de transações no canal loja física	3.72	12.64	15.97	9.83	12.31
Presenças ao longo da época (jogo a jogo)	0.96	1.03	0.06	0.77	0.62
Assiduidade Red Pass	0.06	0.57	0.86	0.06	0.78
Antiguidade de Red Pass	0.27	5.17	9.77	1.34	9.79
Valor total gasto em Red Pass	17.24	239.60	649.45	18.53	1315.97
Presenças em jogos de futebol por mercado secundário, na presente época	0.14	0.14	0.04	0.1	0.14
Número de modalidades assistidas	0.74	1.54	1.54	0.66	1.85
Número total de bilhetes para visitas ao estádio/museu	0.1	0.23	0.16	0.08	0.31
Número de cartões Mais Vantagens ativos	0.18	0.67	0.7	0.34	0.67
Valor total das redenções (saldo utilizado)	2.48	36.91	83.39	7.9	166.24
Longevidade da adesão às Mais Vantagens, em meses	2.23	15.16	16.86	8.92	13.93
Número de vezes que sócio visitou o site	20.08	104.4	80.41	23.37	139.57
Longevidade do registo do sócio no site (meses)	18.81	65.15	61.53	51.51	60.95

Com base nestes indicadores, obtiveram-se as seguintes elações:

- Estes são os sócios com a segunda média de idades mais baixa;
- A última entrada no estádio apresenta um valor médio baixo, muito semelhantes aos grupos de sócios mais relacionados;
- O número de encomendas realizadas é alto e muito próximo do grupo 3 (forte relação com o clube), apesar de ficar um pouco abaixo relativamente aos valores destas transações;
- Através da variável "Diferencial entre valor na presente época e valor médio por época (encomendas)" percebe-se que, de todos, estes foram os sócios que melhores relações com o clube desenvolveram, da época 2021/2022 para a época 2022/2023 (é possível fundamentar esta forte melhoria com base nos bons resultados desportivos da equipa de futebol, na época mais recente);
- Observando ambos os valores médios dos indicadores "Valor total de encomendas realizadas" e "Número de transações em *merchandising*", entende-se que este grupo representa, de facto, um meio termo, para estas matérias, entre os grupos 1 e 4 e os grupos 3 e 5;

- Relativamente a canais de venda, percebe-se que estes sócios são fortes utilizadores dos canais digitais, apresentando valores na ordem dos grupos 3 e 5. Quanto às lojas físicas, e ao contrário dos canais digitais, as transações nestes estabelecimentos não apresentam valores tão acima do expectável, embora o valor médio esteja, novamente, mais próximo dos grupos com relações mais fortes com o clube;
- Este é o grupo com mais presenças em jogos de futebol, através da aquisição de bilhetes jogo a jogo;
- A "Assiduidade Red Pass" foi um dos valores mais interessantes de estudar. O valor "0.57" não é típico para esta variável, uma vez que, geralmente, quando um sócio detém "Red Pass", frequenta a grande maioria dos jogos (ou apresenta o valor "0" se não o tiver adquirido). Este valor pode indicar que neste grupo há uma junção entre sócios com e sem "Red Pass", mas também, pode sugerir que são detentores de "Red Pass" que não sentem a necessidade de marcar presença em absolutamente todos os jogos;
- O valor da "Antiguidade de Red Pass" representa, novamente, um meio termo entre os valores apresentados nos grupos 1 e 4 e nos grupos 3 e 5;
- Quanto ao valor médio da aquisição de "Red Pass", este é relativamente baixo, aproximando-se do comportamento dos grupos com pior relação com o clube. Novamente, este dado pode ser justificado pela combinação entre sócios com e sem "Red Pass", mas também, pode demonstrar o comportamento particular de alguns sócios, com Red Pass, embora um pouco menos afeiçoados à equipa, optando por adquirir "Red Pass's" de valores inferiores, para frequentarem os jogos de forma mais ocasional;
- Quanto às assistências através da aquisição de bilhetes pelo mercado secundário, este é o grupo que apresenta o valor mais alto, a par dos grupos 1 e 5;
- Na variável "Número de modalidades assistidas", o valor médio volta a estar muito próximo dos grupos de sócios fortemente ligados ao clube;
- Quanto à aquisição de bilhetes para visitas ao estádio e ao museu, observa-se um valor significativamente alto, estando mesmo acima do valor do grupo 3;
- Em relação às variáveis respeitantes ao programa "Mais Vantagens", estas oferecem informação algo contraditória. Por um lado, através do número de cartões, entende-se que este é um grupo de sócios onde a maioria são aderentes, mas também através do recurso à longevidade da adesão, que constitui o segundo valor mais alto entre todos os grupos. Por outro lado, no que diz respeito a consumos associados, estes sócios acabam por estar mais próximos dos valores inferiores, como é possível constatar na variável referente ao saldo utilizado. Assim, percebe-se que apesar de uma parte considerável dos sócios ser aderente a esta modalidade, e até com uma antiguidade elevada, estes não têm por hábito realizar transações nos parceiros, nem usufruir do saldo acumulado nas lojas Benfica;

- Relativamente à utilização do site, este é o grupo com os melhores valores. Como observado anteriormente, estes sócios são utilizadores proficientes dos canais digitais, o que também é suportado pelas variáveis referentes ao site. Este grupo tem o segundo maior valor médio de visitas ao site, e o valor médio mais alto da longevidade do registo.

Feita a análise ao grupo 2, considerámos que as conclusões que melhor contribuem para a caracterização deste grupo de sócios são:

→ Recorrentemente, este grupo apresenta valores intermédios face à generalidade;

→ É um grupo composto por sócios com uma média de idades jovem, e com aptidão para os meios digitais;

→ São, em média, os sócios que melhor desenvolveram a sua relação com o clube ao longo do horizonte temporal, ou seja, entre as épocas 2021/2022 e 2022/2023;

→ Com base na observação das variáveis referentes quer a número de transações, quer a valores, para as várias áreas do negócio, depreende-se que estes sócios detêm um menor poder de compra (número considerável de transações, mas valores baixos associados);

→ Apresentam relações moderadas a fortes, com a grande maioria das diversas unidades de negócio.

4.2.5. Grupo 3

Analisando agora os grupos marcados pelas relações fortes entre sócios e clube. Iniciamos a análise, com recurso aos *dashboards*, que desenvolvemos em Power BI. Na Figura 13 constam os gráficos referentes às variáveis demográficas e gerais de consumo.

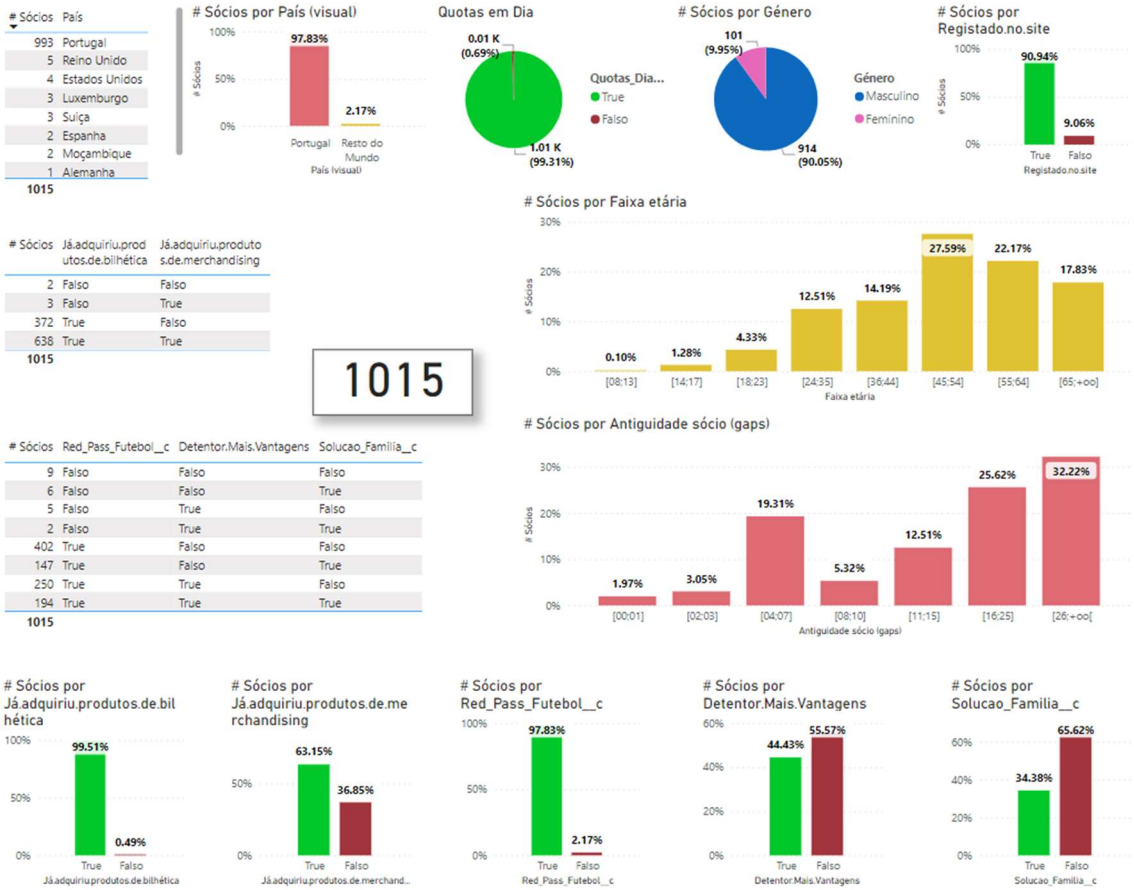

Figura 13 - Indicadores gerais para análise do grupo 3

Pela observação deste *dashboard*, e comparando-o aos restantes, é possível concluir que este é o grupo com maior percentagem de sócios:

- Residentes em Portugal;
- Com os pagamentos de quotas em dia, ou adiantados;
- Com idades entre 45 e 54 anos;
- Que já adquiriram tanto produtos de *merchandising* como produtos de bilhética;
- Aderentes às restantes 3 áreas do negócio representadas (“Red Pass”, “Mais Vantagens” e “Solução Família”).

Com o objetivo de estudar mais detalhadamente os valores relativos aos consumos, foram elaborados os gráficos presentes na Figura 14.

Figura 14 - Indicadores referentes a consumos para análise do grupo 3

Partindo dos dados expostos na Figura 10, concluiu-se que este grupo representa a maior percentagem de sócios:

- Com um valor de quotas pago entre 1,000€ e 2,000€, salientando também que, aproximadamente, 92% das observações deste grupo, estão compreendidas entre os 500€ e 2,000€ de quotas pagas;
- Com 0% de assiduidade através da compra de bilhetes jogo a jogo;
- Com assiduidade "Red Pass" entre os 75% e 100%;
- Que registaram valores de consumo de "Red Pass" entre 500€ e 1,000€;
- Cujos consumos de bilhética jogo a jogo estão entre 50€ e 100€;
- Que acumularam saldo em carteira virtual, com montantes entre 50€ e 150€;
- Aderentes às "Mais Vantagens" há, pelo menos, 48 meses (4 anos).

No âmbito de elaborar mais a análise, foi estudada a tabela de médias e selecionadas variáveis pertinentes para a caracterização dos sócios deste grupo. As variáveis selecionadas estão presentes na Tabela 8.

Tabela 8 - Tabela de médias das variáveis selecionadas para a definição do grupo 3

	C1	C2	C3	C4	C5
Idade	27.99	40.18	50.29	47.41	48.72
Última entrada no estádio	4.46	2.94	1.20	11.13	1.72
Valor total de encomendas realizadas	129.59	745.91	1185.07	291.12	2637.78
Valor de quotas em dívida	8.31	2.07	0.87	3.35	3.13
Longevidade da adesão à "Solução Família", em meses	7.15	7.07	11.38	4.03	7.79
Valor total gasto em merchandising	19.54	84.99	110.94	20.37	257.12
Número de transações no canal site	1.11	4.36	2.53	0.73	9.12
Número de transações no canal APP	0.47	1.49	0.91	0.21	1.54
Número de transações no canal loja física	3.72	12.64	15.97	9.83	12.31
Número de Red Pass's detidos , na presente época	0.07	0.65	1.00	0.06	1.27
Assiduidade Red Pass	0.06	0.57	0.86	0.06	0.78
Somatório do valor acumulado (mercado secundário)	1.56	12.26	23.69	0.99	20.76
Rácio entre número de bilhetes para futebol e número total de bilhetes	0.46	0.85	0.94	0.46	0.89
Diferencial entre consumo na presente época e consumo médio por época (bilhetes de futebol)	10.22	4.39	-28.15	5.02	-18.22
Número de bilhetes para visitas ao estádio/museu, na presente época	0.08	0.14	0.10	0.05	0.19
Número de parceiros onde acumulou saldo em CV	0.79	3.39	4.11	1.52	3.54
Longevidade da adesão às Mais Vantagens, em meses	2.23	15.16	16.86	8.92	13.93
Número de vezes que sócio visitou o site	20.08	104.40	80.41	23.37	139.57
Longevidade do registo do sócio no site (meses)	18.81	65.15	61.53	51.51	60.95

Analisando a tabela, foi possível obter as seguintes conclusões:

- Os sócios deste grupo apresentam a maior média de idades;
- A variável "Última entrada no estádio (meses)" representa o valor mais baixo entre todos os grupos;
- Quanto ao valor médio gasto com consumos gerais, verifica-se um valor consideravelmente elevado;
- Como já observado nos *dashboards*, apresenta o valor médio mais baixo relativamente a pagamentos de quotas em atraso;
- O valor elevado da variável "Longevidade da adesão à "Solução Família", em meses" funciona como um bom indicador para evidenciar a forte relação destes sócios com esta modalidade;
- O valor médio gasto em *merchandising* também é algo elevado;
- Quanto às variáveis que respeitam aos canais de venda, estas apresentam valores relativamente baixos nos meios digitais. Contudo, estes são os sócios que mais transações realizam, em média, nas lojas físicas do Sport Lisboa e Benfica;
- O valor médio de "Red Pass's" detidos na presente época corresponde a 1, o que significa que, aproximadamente, todos os sócios detêm "Red Pass" neste grupo;

- A assiduidade de “Red Pass” apresenta o valor médio mais alto entre todos os grupos;
- Quanto ao somatório do valor acumulado através do mercado secundário, este grupo apresenta o maior valor. Este dado é indicador de que estes sócios dedicam parte do seu tempo a disponibilizar o lugar em mercado secundário, quando não podem estar presentes num determinado jogo. Este comportamento demonstra que os sócios não pretendem ser prejudicados na sua assiduidade, de “Red Pass”, e consequentemente são fortes candidatos a renovar o seu lugar cativo, para a época seguinte;
- O rácio entre bilhetes para futebol e total de bilhetes, representa o maior valor entre os 5 grupos, indicando que, muito provavelmente, estes são os sócios mais ligados à equipa de futebol, porém, poderão também ser os menos interessados nas restantes modalidades;
- Observa-se o valor mais baixo entre os cinco grupos para a variável “Diferencial entre consumo na presente época e consumo médio por época (bilhetes de futebol)”. Este dado indica que estes sócios foram aqueles que mais reduziram o consumo de bilhetes, jogo a jogo, da época 2021/2022 para a época 2022/2023. Esta redução na aquisição de bilhetes não significa, necessariamente, que estes sócios se tenham desligado da equipa, mas pode sugerir que alteraram a maneira como adquiriram os bilhetes. A hipótese mais admissível propõe que a compra dos bilhetes passou da modalidade jogo a jogo para a aquisição de “Red Pass”. Desta forma, depreende-se que este tenha sido o grupo onde o interesse na adesão ao “Red Pass”, cresceu mais significativamente, entre as duas épocas;
- Relativamente a bilhetes para visitas ao estádio e ao museu, estes sócios apresentam um valor médio consideravelmente baixo. Uma forte justificação para este dado, é o facto destes sócios serem aqueles que mais frequentam o estádio, devido à assistência aos jogos, e como tal, acabam por estar menos interessados em adquirir bilhetes para realizar as visitas;
- Através das variáveis referentes à modalidade “Mais Vantagens”, pode verificar-se que estes são os sócios mais relacionados com esta área do negócio. Tanto no número de parceiros como na longevidade da adesão, este grupo apresenta os valores mais altos;
- Tal como visto nos canais de venda, estes sócios não utilizam frequentemente as plataformas digitais do clube. Pelas duas variáveis, destacadas na tabela, referentes ao site, percebe-se que a frequência com que utilizam esta plataforma é baixa, uma vez que, o número de visitas é pouco significativo face ao valor, consideravelmente alto, da longevidade de registo.

Analisados estes dados, foi possível retirar as seguintes conclusões face aos sócios que integram o grupo 3:

- → Numa perspetiva global, estes são os sócios com uma melhor relação com o clube, e os mais comprometidos;
- → Neste grupo estão os sócios que, simultaneamente, têm mais impacto (adesões e consumos) em mais áreas do negócio;
- → É um grupo constituído por sócios com um valor médio de idades mais elevado, o que pode explicar a pouca aptidão para os meios digitais;
- → Estes são os sócios que mais frequentam o estádio, quer seja pelos consumos realizados na loja, quer pela assistência aos jogos, nomeadamente de futebol.

4.2.6. Grupo 5

Finalmente, observamos ao detalhe os dados do grupo 5, para terminar a análise individual dos 5 grupos. Começando, novamente, pelos dados demográficos e gerais, no que respeita aos consumos, através dos *dashboards* desenvolvidos no Power BI, contidos na Figura 15.

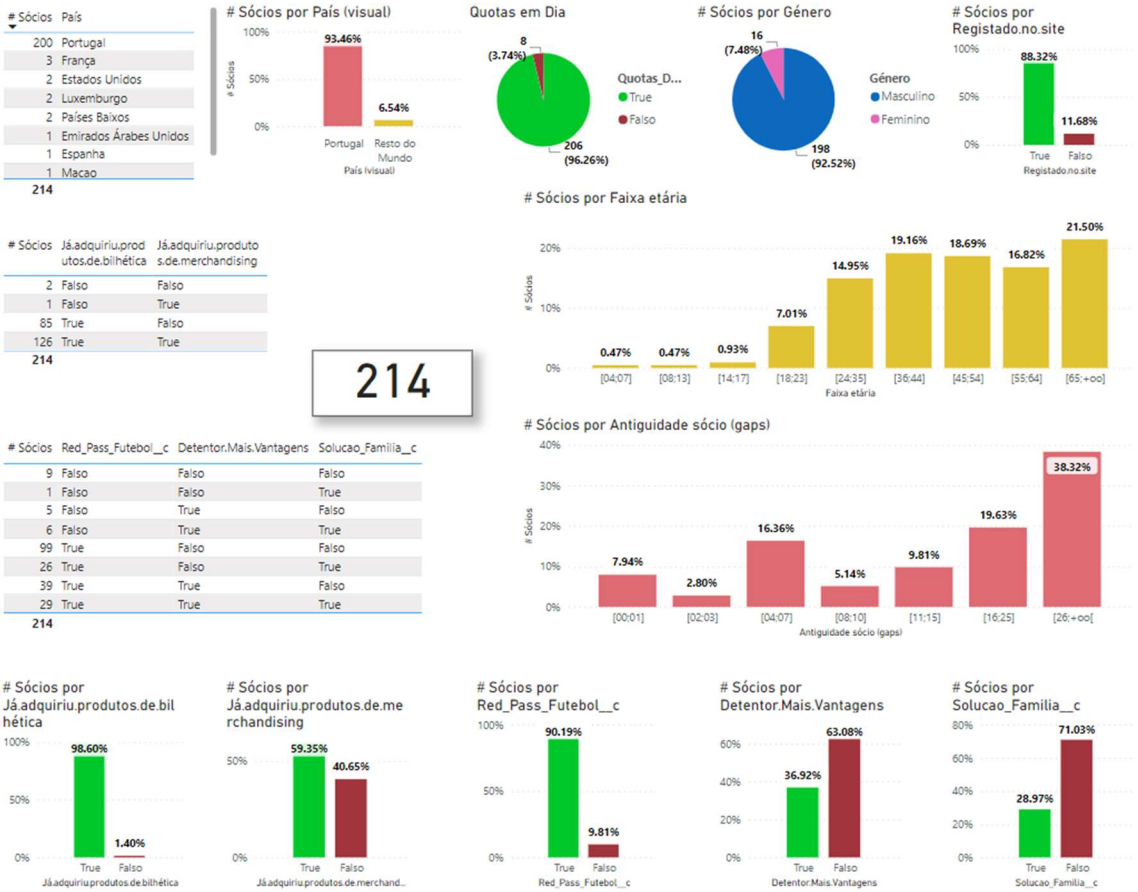

Figura 15 - Indicadores gerais para análise do grupo 5

Estudando os elementos visuais apresentados, e comparando-os com os dos restantes grupos, foi possível concluir que este é o grupo com maior percentagem de sócios:

- Com idades superiores a 65 anos;
- Cuja antiguidade de sócio é igual ou superior a 26 anos.

Prosseguindo a análise com as variáveis que detalham o comportamento dos sócios, face aos consumos nas várias unidades de negócio. Foram elaborados os gráficos da Figura 16 para estudar estas estatísticas.

Figura 16 - Indicadores referentes a consumos para análise do grupo 5

Através da observação destes gráficos, e pela comparação feita com os outros grupos, concluiu-se que este é o grupo com maior percentagem de sócios:

- Com valores de consumos gerais superiores a 1,000€;
- Com valores gastos em "Red Pass" também superiores a 1,000€;
- Com os maiores valores de saldo acumulado em carteira virtual, neste caso, 150€ ou mais;
- Que visitaram o site acima de 150 vezes.

Para concluir a análise deste último grupo, foi elaborada outra tabela de comparação de médias, a Tabela 9, onde foram selecionadas as variáveis mais relevantes para definir os traços comportamentais destes sócios.

Tabela 9 - Tabela de médias das variáveis selecionadas para a definição do grupo 5

	C1	C2	C3	C4	C5
Antiguidade de sócio (anos)	3.44	13.41	20.39	17.49	21.77
Valor total de encomendas realizadas	129.59	745.91	1185.07	291.12	2637.78
Número de casos	0.76	1.77	1.78	0.50	3.08
Antiguidade da dívida, em meses	0.26	-0.58	-0.88	-0.08	-2.49
Número de membros da "Solução Família"	1.64	1.51	1.71	1.27	1.77
Valor total gasto em merchandising	19.54	84.99	110.94	20.37	257.12
Quantidade total de produtos adquiridos (merchandising)	1.52	6.97	8.40	1.86	22.73
Número de transações no canal site	1.11	4.36	2.53	0.73	9.12
Número de transações no canal APP	0.47	1.49	0.91	0.21	1.54
Número de transações no canal phygital	0.01	0.03	0.03	0.01	0.22
Assiduidade Red Pass	0.06	0.57	0.86	0.06	0.78
Número de vezes que o lugar de Red Pass foi utilizado por alguém que não o sócio, na presente época	0.24	2.09	2.54	0.21	3.37
Valor total gasto em Red Pass	17.24	239.60	649.45	18.53	1315.97
Número de bilhetes adquiridos para modalidades	0.45	2.07	3.37	0.39	9.44
Número de bilhetes para visitas ao estádio/museu	0.10	0.23	0.16	0.08	0.31
Número total de bilhetes adquiridos	4.75	33.02	50.08	4.72	74.42
Número de redenções	0.15	1.52	2.19	0.31	3.25
Saldo acumulado	6.60	67.04	144.23	18.48	230.62
Longevidade da adesão às Mais Vantagens, em meses	2.23	15.16	16.86	8.92	13.93
Número de vezes que sócio visitou o site	20.08	104.40	80.41	23.37	139.57
Longevidade do registo no site	18.81	65.15	61.53	51.51	60.95

Através da observação desta tabela, foi possível retirar as seguintes conclusões, em relação a este grupo de sócios:

- Este é o grupo que apresenta o maior valor mais elevado na variável "Antiguidade de sócio (anos)";
- Verifica-se que este é o grupo com o valor mais alto de encomendas realizadas, tendo uma margem muito significativa quanto aos restantes;
- O número de casos é alto face aos restantes grupos, demonstrando a facilidade com que estes sócios comunicam com o clube;
- A antiguidade da dívida representa o valor negativo mais baixo, o que indica que estes são os sócios com um maior valor de quotas pagas em adiantado;
- Quanto à modalidade da "Solução Família", em média, neste grupo, estão os sócios com mais membros (também eles sócios) agregados à sua família;
- Em relação à aquisição de produtos *merchandising*, tanto o valor total gasto como a quantidade de produtos adquiridos, demonstram que neste grupo estão os sócios mais consumistas. Da mesma forma, observando as variáveis referentes aos consumos pelos

canais de venda, percebe-se que são os sócios com mais transações nos canais site, app e phygital;

- A assiduidade de "Red Pass" tem um valor elevado, demonstra que praticamente todos os sócios deste "cluster" detêm "Red Pass" e frequentem a maioria dos jogos;
- Por outro lado, o valor da assiduidade também é impactado pela utilização do mercado secundário. Como se pode observar na tabela, estes são os sócios que colocam bilhetes em mercado secundário mais frequentemente;
- Para concluir a análise aos indicadores do "Red Pass", verifica-se que estes foram os sócios que gastaram valores mais expressivos, com a aquisição dos lugares mais caros;
- Pela compra de bilhetes para jogos das restantes modalidades, percebe-se que este é o grupo de sócios mais envolvidos com os vários desportos praticados pelo clube, além do futebol;
- Quanto às visitas ao estádio e ao museu, este grupo apresenta os valores mais altos. Este dado foi considerado inconclusivo e contraditório, uma vez que praticamente todos os sócios frequentam regularmente o estádio, devido à detenção de "Red Pass", e como tal, seria de esperar que o interesse em adquirir bilhetes para visitas fosse menor;
- Para concluir a análise às variáveis de bilhética, verificou-se que este é o grupo com mais bilhetes adquiridos, sendo um valor bastante significativo, quando comparado com os outros grupos. Este dado indica-nos que grande parte dos "Red Pass's" destes sócios proporcionam o acesso a todas as competições, garantindo-lhes assim mais bilhetes;
- Nas variáveis referentes às "Mais Vantagens", apesar de valores elevados, não são o grupo superlativo em todas elas. Na questão da longevidade, representam apenas o terceiro valor mais alto entre os 5 grupos, porém, mais uma vez, em matéria de consumos excedem todos os outros. Um dado interessante é o de que embora não apresentem, em média, uma longevidade de adesão às "Mais Vantagens" tão elevada como outros grupos, os sócios do grupo 5, detêm um saldo acumulado muito superior aos restantes. Isto demonstra que realizam mais transações ou consumos mais dispendiosos, nos parceiros, em relação aos sócios dos outros grupos;
- Por fim, à semelhança dos indicadores do programa "Mais Vantagens", verifica-se que, apesar de não serem os sócios com o maior valor da "Longevidade do registo no site", têm muito mais visitas, em média, que os restantes grupos.

Finalizada a análise, estas são as principais conclusões a reter sobre o comportamento destes sócios:

- → Este grupo é constituído por sócios com, em média, os valores mais elevados quanto à antiguidade de adesão, fazendo deles, os mais comprometidos com o clube;
- → Neste grupo estão os sócios mais consumistas, ou com maior poder de compra, apresentando valores superlativos e de grande destaque, nas variáveis de consumo, em praticamente todas as áreas do negócio;
- → Não apresentam os melhores valores em absolutamente todas as áreas do negócio, mas, em geral, têm uma relação forte com todas elas.

4.2.7. Principais diferenças entre os grupos 3 e 5

Após conhecer melhor, individualmente, as características dos sócios presentes nos grupos 3 e 5, o objetivo passa por compará-los e entender aquilo que os distingue. Para tal, foram selecionadas as variáveis presentes na Tabela 10, consideradas as mais adequadas para observar estas diferenças.

Tabela 10 - Tabela de médias das variáveis selecionadas para a comparação entre os grupos 3 e 5

	C1	C2	C3	C4	C5
Última entra no estádio	4.46	2.94	1.20	11.13	1.72
Valor total das encomendas realizadas	129.59	745.91	1185.07	291.12	2637.78
Valor de quotas em dívida	8.31	2.07	0.87	3.35	3.13
Número de membros da "Solução Família"	1.64	1.51	1.71	1.27	1.77
Longevidade da adesão à "Solução Família", em meses	7.15	7.07	11.38	4.03	7.79
Valor total gasto em merchandising	19.54	84.99	110.94	20.37	257.12
Número de transações no canal site	1.11	4.36	2.53	0.73	9.12
Número de transações no canal APP	0.47	1.49	0.91	0.21	1.54
Número de transações no canal loja física	3.72	12.64	15.97	9.83	12.31
Assiduidade Red Pass	0.06	0.57	0.86	0.06	0.78
Valor total gasto em Red Pass	17.24	239.60	649.45	18.53	1315.97
Número de bilhetes adquiridos para modalidades	0.45	2.07	3.37	0.39	9.44
Número total de bilhetes adquiridos	4.75	33.02	50.08	4.72	74.42
Saldo acumulado	6.60	67.04	144.23	18.48	230.62
Longevidade da adesão às Mais Vantagens, em meses	2.23	15.16	16.86	8.92	13.93
Número de vezes que sócio visitou o site	20.08	104.40	80.41	23.37	139.57
Longevidade do registo no site	18.81	65.15	61.53	51.51	60.95

Colmatando as análises individuais com a observação desta tabela, foi possível obter as seguintes conclusões:

→ Pelas variáveis "Última entrada no estádio" e "Assiduidade Red Pass", conseguimos perceber que os sócios do grupo 3 são mais ligados à equipa de futebol;

→ Através dos consumos nos 3 canais de venda e do número de visitas ao site, depreende-se que enquanto os sócios do grupo 3 não são utilizadores tão capazes dos meios digitais, optando por frequentarem as lojas físicas, no grupo 5, os sócios desempenham o comportamento oposto;

→ No grupo 3, regista-se um comportamento dos sócios muito ligado às diferentes unidades do negócio, principalmente, quando se trata da antiguidade de adesão às várias modalidades que o clube oferece. Apesar do grupo 5 não registar uma antiguidade tão alta, estes sócios parecem usufruir mais destes serviços, que os sócios do grupo 3. É possível verificar estes dados observando as estatísticas de antiguidade, mas também as variáveis "Número de membros da solução família", "Saldo acumulado" e "Número de vezes que sócio visitou o site";

→ Quanto aos valores dos consumos, realçamos os valores, muito superiores, do grupo 5. Embora os sócios do grupo 3 tenham valores consideráveis, relativamente aos consumos nas várias áreas do negócio, os valores do grupo 5 são sempre significativamente

superiores. Estes números são apoiados pelas variáveis "Valor total das encomendas realizadas", "Valor total gasto em *merchandising*", "Valor total gasto em Red Pass", "Número total de bilhetes adquiridos" e "Saldo acumulado";

→ Para concluir a informação relativa ao departamento de bilhética, destacar que no grupo 5, os sócios têm muito mais interesse em adquirir bilhetes para as modalidades, além do futebol, que os sócios do grupo 3. A diferença de valores é, mais uma vez, bastante significativa;

→ Para terminar, observa-se através da variável "Valor de quotas em dia" que os sócios do grupo 3 têm, em média, um montante de dívida menor que os sócios do grupo 5.

5. Conclusões

Este trabalho teve como principal objetivo conhecer melhor os perfis de sócios do Sport Lisboa e Benfica. Para tal, utilizámos as bases de dados disponibilizadas pela empresa para realizar uma análise de grupos e classificar o comportamento dos sócios. Além disso, a intenção por detrás do projeto passa por alavancar um novo e mais avançado nível de personalização feito nas campanhas e comunicação aos sócios, uma questão bastante explorada atualmente, na área do CRM. Esta meta é complexa e deve ser trabalhada de forma gradual, não sustentando a tomada de decisões com base em apenas um estudo. Ainda assim, com este trabalho, esperamos estar cada vez mais perto do grau de personalização que é ambicionado, produzindo um novo *output* com uma visão e respostas diferentes, sobre esta temática.

Nos resultados do projeto percebe-se que a quantidade de sócios mais afastados do clube é muito superior aos mais ativos. Na amostra utilizada, mais de 75% das observações correspondem a sócios com uma ligação mais fraca ao Sport Lisboa e Benfica. Além de serem números bastante significativos, não são favoráveis para o negócio. No algoritmo de agrupamento foram criados dois grupos para dividir estas observações, com o intuito de melhorar a personalização, e não generalizar os sócios menos ativos para um único grupo. Com esta decisão foi possível encontrar os diferentes aspetos do negócio com os quais os sócios dos dois grupos mais se relacionam, promovendo estratégias diferentes para voltar a cativar estes sócios, de uma maneira mais apelativa e personalizada. Desta forma, estamos também a contribuir para outro grande objetivo do CRM: a recuperação dos sócios em *churn*.

Além do valor que este estudo acrescenta ao CRM do Sport Lisboa e Benfica, deve ser ainda mencionado o trabalho a realizar daqui para a frente, na sequência do projeto. Olhando para as limitações encontradas ao longo deste trabalho, é de realçar que a qualidade dos dados em "Benficómetro" foi, por vezes, um obstáculo. Outra dificuldade sentida foi no desenvolvimento de variáveis mais complexas, nomeadamente referentes a frequências, ou modas, de consumos (por exemplo, qual a família de produtos mais comercializada pelo sócio), devido à forma como os dados estão estruturados em "Benficómetro". Houve tentativas de desenvolver estas variáveis em Power BI, mas sem sucesso, tendo sido retiradas do modelo. Para além das dificuldades, na sequência deste trabalho, seria interessante pôr em prática campanhas personalizadas, com base nos perfis designados. Posteriormente, os resultados das campanhas poderão ser analisados, de modo a perceber se o modelo desenvolvido neste trabalho teve qualidade e utilidade, na ótica da empresa.

Referências Bibliográficas

Abdi, F., & Abolmakarem, S. (2019). Customer Behavior Mining Framework (CBMF) using clustering and classification techniques. *Journal of Industrial Engineering International*, *15*, 1–18. https://doi.org/10.1007/s40092-018-0285-3

Abdullah Al-Suraihi, W., Abdullah Al-Suraihi, A.-H., Ibrahim, I., Al-Tahitah, A., & Abdulrab, M. (2020). The Effect of Customer Relationship Management on Consumer Behavior: A Case of Retail Industry in Malaysia. *International Journal of Management and Human Science (IJMHS)*, *4*(3), 32–40.

Abirami, M., & Pattabiraman, V. (2016). Data mining approach for intelligent customer behavior analysis for a retail store. *Smart Innovation, Systems and Technologies*, *49*, 283–291. https://doi.org/10.1007/978-3-319-30348-2_23

Ackerman, M., & Ben-David, S. (2016). A Characterization of Linkage-Based Hierarchical Clustering. *Journal of Machine Learning Research*, *17*, 1–17.

Ackerman, M., Ben-David, S., & Loker, D. (2010). Towards Property-Based Classification of Clustering Paradigms. *Advances in Neural Information Processing Systems*, *23*.

Aggarwal, C. C., & Yu, P. S. (1999). Data Mining Techniques for Associations, Clustering and Classification. *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 13–23.

Ahmed, S., & Aissa, H. (2018). The impact of customer relationship management on customer behavior: case study of ooredoo for telecommunications. *Revue Des Sciences Commerciales*, *17*. https://ssrn.com/abstract=3192717

Anshari, M., Almunawar, M. N., Lim, S. A., & Al-Mudimigh, A. (2019). Customer relationship management and big data enabled: Personalization & customization of services. *Applied Computing and Informatics*, *15*(2), 94–101. https://doi.org/10.1016/j.aci.2018.05.004

Bair, E. (2013). Semi-supervised clustering methods. *Wiley Interdisciplinary Reviews: Computational Statistics*, *5*(5), 349–361. https://doi.org/10.1002/wics.1270

Berry, L. L., Shostack, G. L., Upah, G. D., & American Marketing Association. (1983). *Emerging perspectives on services marketing*. 25–28.

Bezdek, J. C., Hathaway, R. J., & Huband, J. M. (2007). Visual assessment of clustering tendency for rectangular dissimilarity matrices. *IEEE Transactions on Fuzzy Systems*, *15*(5), 890–903. https://doi.org/10.1109/TFUZZ.2006.889956

Bose, I., & Chen, X. (2015). Detecting the migration of mobile service customers using fuzzy clustering. *Information and Management*, *52*(2), 227–238. https://doi.org/10.1016/j.im.2014.11.001

Chan, I. C. C., Fong, D. K. C., Law, R., & Fong, L. H. N. (2018). State-of-the-art social customer relationship management. *Asia Pacific Journal of Tourism Research*, *23*(5), 423–436. https://doi.org/10.1080/10941665.2018.1466813

Erevelles, N. (2015). Disability Studies as Insight: Deploying Enabling Pedagogies in HIV/AIDS Education. *Humanizing Pedagogy Through HIV and AIDS Prevention*, 31–49.

Feng, C. C., Wang, Y. C., & Chen, C. Y. (2014). Combining Geo-SOM and hierarchical clustering to explore geospatial data. *Transactions in GIS*, *18*(1), 125–146. https://doi.org/10.1111/tgis.12025

Guo, S., Zhao, H., & Yang, W. (2021). Hierarchical feature selection with multi-granularity clustering structure. *Information Sciences*, *568*, 448–462. https://doi.org/10.1016/j.ins.2021.04.046

Hadden, J., Tiwari, A., Roy, R., & Ruta, D. (2007). Computer Assisted Customer Churn Management: State-Of-The-Art and Future Trends. *Computers & Operations Research*, *34*(10), 2902–2917.

Haque, S., Eberhart, Z., Bansal, A., & McMillan, C. (2022). Semantic Similarity Metrics for Evaluating Source Code Summarization. *IEEE International Conference on Program Comprehension*, *2022-March*, 36–47.

Hofacker, C. F., Malthouse, E. C., & Sultan, F. (2016). Big Data and consumer behavior: imminent opportunities. *Journal of Consumer Marketing*, *33*(2), 89–97. https://doi.org/10.1108/JCM-04-2015-1399

Hsu, F. M., Lu, L. P., & Lin, C. M. (2012). Segmenting customers by transaction data with concept hierarchy. *Expert Systems with Applications*, *39*(6), 6221–6228. https://doi.org/10.1016/j.eswa.2011.12.005

Jobson, J. D. (2012). *Applied multivariate data analysis: volume II: Categorical and Multivariate Methods*. Springer Science & Business Media.

Kanavos, A., Iakovou, S. A., Sioutas, S., & Tampakas, V. (2018). Large scale product recommendation of supermarket ware based on customer behaviour analysis. *Big Data and Cognitive Computing*, *2*(2), 1–19. https://doi.org/10.3390/bdcc2020011

Kaya Gülağız, F., & Şahin, S. (2017). Comparison of Hierarchical and Non-Hierarchical Clustering Algorithms. *International Journal of Computer Engineering and Information Technology*, *9*(1), 6–14.

Khade, A. A. (2016). Performing Customer Behavior Analysis using Big Data Analytics. *Procedia Computer Science*, *79*, 986–992. https://doi.org/10.1016/j.procs.2016.03.125

Kim, M. K., Park, M. C., & Jeong, D. H. (2004). The effects of customer satisfaction and switching barrier on customer loyalty in Korean mobile telecommunication services. *Telecommunications Policy*, *28*(2), 145–159. https://doi.org/10.1016/j.telpol.2003.12.003

Kuo, R. J., Ho, L. M., & Hu, C. M. (2002). Cluster analysis in industrial market segmentation through artificial neural network. *Computers & Industrial Engineering*, *42*(2–4), 391–399

Leventhal, B. (2018). *Predictive Analytics for Marketers: Using Data Mining for Business Advantage*. Kogan Page Publishers.

Li, J., Pan, S., Huang, L., & Zhu, X. (2019). A machine learning based method for customer behavior prediction. *Tehnicki Vjesnik*, *26*(6), 1670–1676. https://doi.org/10.17559/TV-20190603165825

Mccalla, G., Greer, J., And, B. B., & Pospisil, P. (1992). Granularity Hierarchies. *Computers & Mathmatics with Applications*, *23*(2–5), 363–375.

Miah, S. J., Vu, H. Q., Gammack, J., & McGrath, M. (2017). A Big Data Analytics Method for Tourist Behaviour Analysis. *Information and Management*, *54*(6), 771–785. https://doi.org/10.1016/j.im.2016.11.011

Mohammadi Nasrabadi, A., Hosseinpour, M. H., & Ebrahimnejad, S. (2013). Strategy-aligned fuzzy approach for market segment evaluation and selection: a modular decision support system by dynamic network process (DNP). *Journal of Industrial Engineering International*, *9*(1). https://doi.org/10.1186/2251-712X-9-11

Murtagh, F., & Contreras, P. (2012). Algorithms for hierarchical clustering: An overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *2*(1), 86–97. https://doi.org/10.1002/widm.53

Murtagh, F., & Legendre, P. (2014). Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? *Journal of Classification*, *31*(3), 274–295. https://doi.org/10.1007/s00357-014-9161-z

Nauck, D. (2013). Predictive Analytics and Proactive Service. *IET Seminar on Data Analytics 2013: Deriving Intelligence and Value from Big Data*, 1–26.

Posit team. (2023). *RStudio: Integrated Development Environment for R*. Posit Software, PBC, Boston, MA.

Provost, F., & Fawcett, T. (2013). Data Science and its Relationship to Big Data and Data-Driven Decision Making. *Big Data*, *1*(1), 51–59. https://doi.org/10.1089/big.2013.1508

Savaş, S., Topaloğlu, N., & Yilmaz, M. (2012). VERİ MADENCİLİĞİ VE TÜRKİYE'DEKİ UYGULAMA ÖRNEKLERİ. *İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi*, *11*(21), 1–23.

Schonlau, M. (2004). Visualizing non-hierarchical and hierarchical cluster analyses with clustergrams. *Computational Statistics*, *19*, 95–111.

Sharma, S. (1995). *Applied multivariate techniques*. John Wiley & Sons, Inc.

Surendro, K. (2019). Predictive analytics for predicting customer behavior. *2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT)*, 230–233.