UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

Estudo da vinculação de um cliente particular a uma plataforma digital

Joana Bilro Banza

Mestrado em Matemática Aplicada à Economia e Gestão

Trabalho de projeto orientado por:
Raquel João Fonseca
João Telhada

2019

Resumo

O presente trabalho é composto por cinco capítulos. No primeiro será feita uma introdução de modo a contextualizar o tema assim como realizar uma breve apresentação da evolução da plataforma digital desde a sua implementação na empresa. Segue-se a Análise de *Clusters* com vista a obter o perfil dos clientes digitais com base nas suas características mais relevantes. No capítulo seguinte será feita uma análise exploratória dos dados, com o objetivo de dar a conhecer as várias operações disponíveis na plataforma digital em estudo e de apresentar uma visão geral do ponto de situação da empresa relativamente aos seus clientes e contratos em vigor. O quarto capítulo será dedicado à análise de sazonalidade. Para identificar a presença de sazonalidade caso exista, será efetuada uma análise de variância seguida de testes *pairwise*, sendo que para isso serão aplicados os testes *Fisher's Least Significant Difference* e *Tukey's Honestly Significant Difference*. Por fim será apresentada a conclusão onde será realizada uma reflexão dos resultados obtidos ao longo do trabalho, assim como sugestões para trabalhos futuros.

Palavras-chave: Análise de *clusters*, *k-means*, *pairwise*, sazonalidade

Abstract

The present work consists of five chapters. In the first one, an introduction will be made to contextualize the theme as well as give a brief presentation of the evolution of the digital platform since it's implementation in the company. This is followed by a Cluster Analysis to get the profile of digital customers based on their most relevant characteristics. In the next chapter an exploratory analysis of the data will be made, in order to know the various operations available on the digital platform under study and to present an overview of the company's situation with respect to it's customers and existing contracts. The fourth chapter will be devoted to seasonality analysis. To identify the presence of seasonality, if any, a variance analysis will be performed followed by pairwise tests, and the Fisher's Least Significant Difference and Tukey's Honestly Significant Difference tests will be applied. Finally will be presented the conclusion where will be a reflection of the results obtained throughout the work, as well as suggestions for future work.

Keywords: Cluster Analysis, k-means, pairwise, seasonality

Agradecimentos

Este projeto representa o final de mais uma etapa da minha vida académica e o começo de outras tão ou mais desafiantes. Nada disto teria sido possível sem o apoio incondicional da minha família, o qual quero agradecer em especial aos meus pais, por nunca me deixarem desistir dos meus objetivos.

Aos meus orientadores, professor João Telhada e professora Raquel Fonseca, por toda a disponibilidade, pelo conhecimento que me transmitiram e por terem sido incansáveis desde o início até ao final deste trabalho.

Queria ainda agradecer à empresa e à minha equipa de trabalho que me acompanhou ao longo destes meses, pela ajuda e partilha de ideias.

Por último, não posso deixar de agradecer aos meus amigos por terem sempre uma palavra de apoio a dizer e por estarem sempre prontos para me dar a mão.

Acredito realmente que a aprendizagem deverá ser sempre uma constante na minha vida e que todo o conhecimento que possa adquirir nunca será demais.

"Somos o resultado dos livros que lemos, dos cafés que desfrutamos, das viagens que fazemos e das pessoas que amamos."

Airton Ortiz

Índice

Lista de Figuras **ix**

Lista de Tabelas **xi**

1 Introdução **1**

2 Análise de *Clusters* **5**
2.1 Método de *k-means* 6
2.2 Análise de resultados 8

3 Análise Exploratória dos Dados **13**
3.1 Clientes 13
3.1.1 Ponto de situação de clientes em carteira a dezembro 2018 14
3.1.2 Visão anual de clientes 15
3.2 Contratos 16
3.2.1 Ponto de situação de contratos em carteira a dezembro 2018 17
3.2.2 Visão anual de contratos 18
3.3 Operações 18
3.3.1 *Downloads* da aplicação 19
3.3.2 *Logins* 20
3.3.3 Comunicação de leituras 21
3.3.4 Consultas 22
3.3.5 Pedidos de Informação 23
3.3.6 Reclamações 24
3.4 Digital *vs* Outros canais 24
3.4.1 Comunicação de Leituras 25
3.4.2 Pedido de Referência Expresso 25
3.4.3 Adesão à Fatura Eletrónica 26
3.4.4 Adesão ao Débito Direto 27

4 Análise de Sazonalidade **29**
4.1 Análise da Variância Simples (*One Way* ANOVA) 29
4.1.1 Partição da Soma de Quadrados 30
4.2 Testes *pairwise* 32
4.2.1 *Fisher's Least Significant Difference* (LSD) 32
4.2.2 *Tukey's Honestly Significant Difference* (TSD) 33
4.3 Resultados 34

4.3.1 *Downloads* 34
4.3.2 *Logins* 36
4.3.3 Consumos 37
4.3.4 Reclamações 40

5 Conclusão **43**

Lista de Figuras

2.1 Exemplo do número ótimo de *clusters* 7
2.2 Número ótimo de *clusters* 8
2.3 Número ótimo de *clusters* 9

3.1 Clientes na carteira a 31 de dezembro de 2018 15
3.2 Visão geral de clientes (k) 16
3.3 Contratos na carteira a 31 de dezembro de 2018 17
3.4 Visão geral de contratos (k) 18
3.5 Número de downloads da aplicação (unidades) 20
3.6 Número de logins (k) 21
3.7 Número de comunicação de leituras (k) 22
3.8 Número de consultas (k) 23
3.9 Número de pedidos de informação (k) 23
3.10 Número de reclamações 24
3.11 Comunicações de leitura na plataforma digital *vs* noutros canais 25
3.12 Pedidos de referência expresso na plataforma digital *vs* noutros canais 26
3.13 Adesões à fatura eletrónica na plataforma digital *vs* noutros canais 27
3.14 Adesões ao débito direto na plataforma digital *vs* noutros canais 28

Lista de Tabelas

2.1 Tipo de consumo por *cluster* 10
2.2 Idade por *cluster* 10
2.3 Região por *cluster* 10
2.4 Volume de clientes digitais por *cluster* 11

4.1 ANOVA 31
4.2 Pares a comparar 32
4.3 Número de *downloads* da aplicação por trimestre 34
4.4 ANOVA - *downloads* 34
4.5 Diferenças entre as médias das comparações *pairwise* - *downloads* 35
4.6 Número de *logins* na plataforma digital por trimestre 37
4.7 ANOVA para *logins* 37
4.8 Consumos médios por trimestre 37
4.9 ANOVA para consumos 38
4.10 Diferenças entre as médias das comparações *pairwise* - consumo 39
4.11 Número de reclamações efetuadas por trimestre 40
4.12 ANOVA para reclamações 40
4.13 Diferenças entre as médias das comparações *pairwise* - reclamações 41

Capítulo 1

Introdução

Facilmente reconhecemos que, ao longo dos últimos anos, o número de plataformas digitais tem vindo a desenvolver um aumento bastante notório a nível global. Estas são canais de interação da empresa com os consumidores e são concebidas com o objetivo de que os mesmos se possam conectar e interagir entre si de modo a criar valor, proporcionando assim uma maior proximidade entre a empresa e o cliente. Esta tecnologia veio revolucionar a estratégia de mercado das empresas com o objetivo de melhorar os resultados e otimizar processos, sendo esta estratégia variável de empresa para empresa e desenvolvida conforme o que fizer mais sentido para a organização e aqueles que são os seus objetivos a longo prazo.

Neste tipo de plataforma deve ter-se sempre presente que o foco principal é melhorar a experiência do utilizador, com vista a reter os seus clientes atuais assim como atrair novos, até porque o sucesso desta dependerá da forma como os seus utilizadores se adaptam a ela. Neste sentido, surge o conceito de *Customer Experience* (CX) que tem também ele vindo a crescer e a ganhar o seu lugar nos últimos anos, sendo cada vez mais uma das maiores preocupações das empresas. *Customer Experience* consiste essencialmente na perceção que o cliente tem a respeito da empresa, o que torna muito importante que cada momento de interação entre ambos seja o mais satisfatório possível, de modo a garantir que não se perde a atenção do cliente e ainda que exista uma recomendação da plataforma a potenciais utilizadores. Este fenómeno surge num momento em que as empresas começam cada vez mais a compreender que, por todas estas razões, o cliente deve ser o centro de todas as decisões, uma vez que só assim existirá uma relação de lealdade entre eles de modo a fortalecer o negócio. É extremamente importante ter em mente que este conceito é muito frágil no sentido em que a perceção do utilizador em relação à plataforma pode variar a cada interação com a mesma, sendo assim essencial que se mantenha uma ligação sempre constante ao longo do ciclo de vida da relação com o cliente.

Com a vasta oferta de mercado, torna-se cada vez mais complicado para uma empresa reter o cliente, tendo para isso de se ser o mais inovador e apelativo possível nas ofertas ao consumidor. Para acompanhar toda esta mudança, torna-se indispensável criar novas metodologias de trabalho e deixar de parte conceitos demasiado conservadores.

Este trabalho está relacionado com a plataforma digital de uma empresa no setor de energia e tem como objetivo compreender de que forma pode um cliente estar vinculado a ela, tendo em conta o seu perfil enquanto utilizador da mesma.

Para que isto seja possível, a área funcional da empresa que faz a gestão dos canais digitais torna-se imprescindível no sentido de promover o estudo de clientes digitais, de modo a conhecer o seu perfil, as suas preferências e as suas necessidades. Após esta análise, é possível gerar campanhas personalizadas

para cada tipo de utilizador, sem que haja uma generalização que tenha impacto em clientes com necessidades diferentes. Deste modo, a empresa consegue promover a área de cliente e angariar clientes digitais, assim como verificar se as medidas implementadas estão a produzir os resultados pretendidos e, em caso negativo, propor medidas corretivas. Entende-se por cliente digital aquele que adere à utilização de plataformas digitais que lhe permitam interagir com a empresa de modo remoto evitando uma interação presencial e, consequentemente, mais demorada. Nesta plataforma pode então ser gerida toda a relação com a empresa em qualquer altura e em qualquer lugar, com a máxima comodidade.

A área reservada da plataforma em estudo foi disponibilizada no ano de 2013 focando-se, inicialmente, em clientes particulares e disponibilizando funcionalidades básicas como alterações contratuais e envio de contactos. Até 2015 ocorreu uma evolução de modo a abranger clientes empresariais, permitindo também novas funcionalidades tais como comunicação de leituras e comparação de consumos. Acompanhando este avanço tecnológico foi também neste ano que foi publicada a aplicação com o propósito de tornar mais apelativa a interação do utilizador com a plataforma assim como a atingir um público alvo cada vez mais ligado ao digital.

A fase seguinte (2016-2017) focou-se em melhorar a experiência do cliente e em resolver pontos críticos como erros de utilização e introdução de utilidades até então não disponíveis. Neste processo, destacaram-se iniciativas como a migração da plataforma para uma nova tecnologia, a adoção de um novo *layout* com o objetivo de tornar a navegação do cliente mais rápida e simples, a criação de um registo simplificado e a reformulação da área de apoio ao cliente.

A área de cliente agrega toda a informação relativa às casas e/ou negócios dos clientes, ou ainda das casas ou negócios que estes se encontram a gerir. Se o cliente tiver mais do que um contrato, é-lhe ainda permitido gerir a sua carteira de contratos de uma forma simples e rápida, através da criação de grupos ou mesmo personalizando o seu contrato. Estas funcionalidades tornam muito mais fácil e intuitiva a interação do utilizador com a empresa, evitando assim deslocações desnecessárias e menos cómodas para o cliente. No entanto, é ainda muito significativa a quantidade de clientes que opta pelos canais assistidos para efetuar interações com a empresa. Apesar deste dito avanço no mundo do digital, existe ainda alguma resistência às novas tecnologias por parte de alguns consumidores, seja pela sua faixa etária ou por mera preferência dos meios tradicionais.

Para que os utilizadores tenham acesso à plataforma em estudo deverão passar por um processo de adesão, após o qual o utilizador estará apto a fazer as suas operações através da mesma. O que acontece muitas vezes é que, após o registo, os clientes não chegam a realizar a ativação do mesmo e, por esse motivo, não poderão aceder à plataforma de modo a gerir os seus contratos de energia.

A frequência com que os utilizadores acedem à plataforma e executam as suas operações, isto é, o seu comportamento digital, ditará se estes estão ou não vinculados a ela. Será então realizado um estudo da vinculação do cliente à plataforma como sendo a ligação que estes estabelecem entre si, ou seja, se não só o utilizador está satisfeito com ela como também continua a preferir continuamente este serviço face a outros existentes no mercado. É importante compreender que a satisfação do utilizador pode ser algo momentâneo e para que se considere a existência de vinculação, a ligação entre o cliente e a plataforma deve ser algo a longo prazo. Isto porque com o avanço tecnológico vem também todo um leque de cada vez mais opções e a oferta aumenta significativamente. É, por isso, cada vez mais difícil e desafiante prender a atenção dos consumidores.

Certamente que, dada a natureza da plataforma, é expectável que possam existir alguns efeitos de sazonalidade. A existirem, é importante para a empresa perceber de que forma se poderá precaver esta

situação de modo a manter a plataforma digital rentável. A presença de sazonalidade é bastante comum neste tipo de negócio, o que fará diferença será a forma como a empresa se protege dos efeitos negativos e como aproveita os positivos para alavancar os momentos menos bons. Tratando-se de uma empresa no setor de energia, à primeira vista é evidente que existirá uma sazonalidade pelo menos ao nível do consumo, dado as alterações climáticas que acontecem ao longo do ano. O desafio será perceber em que medida é que isso poderá impactar a forma como os utilizadores interagem com a plataforma digital.

Assim, realizou-se uma análise de *clusters* utilizando o método de *k-means*, com o objetivo de agrupar os clientes com base no seu consumo e faixa etária, de modo a perceber a relação entre os dois. Para que isto fosse possível, foi utilizado o programa *RStudio*, que permitiu aplicar o algoritmo referido. Seguidamente, são dadas a conhecer as várias operações disponíveis na plataforma em estudo, assim como os resultados obtidos ao longo do ano, para que se compreenda melhor o funcionamento da mesma. O quarto capítulo tem como objetivo determinar a presença de sazonalidade, caso exista, através da análise de variância e de testes *pairwise*, sendo para isso aplicados os testes *Fisher's Least Significant Difference* e *Tukey's Honestly Significant Difference*. Por último, é feita uma reflexão dos resultados obtidos ao longo do trabalho.

Este trabalho tem como objetivos principais definir o perfil do cliente digital com base nas suas características assim como estudar o seu comportamento digital.

Capítulo 2

Análise de *Clusters*

A análise de *clusters* é bastante interessante do ponto de vista do *marketing* estratégico da empresa na medida em que permite uma comunicação mais individualizada. Deste modo, torna-se mais fácil compreender quais são os grupos de pessoas mais recetivos a determinadas campanhas e posterior adesão a serviços.

De modo geral, este tipo de análise consiste no processo de criação de grupos, ou *clusters*, partindo de uma amostra com n observações. A cada iteração, os dados são distribuídos de maneira a que observações semelhantes se encontrem no mesmo grupo enquanto que observações distintas se encontram em grupos diferentes.

Esta análise pode ser feita com base em dois métodos diferentes (Tan et al., 2006):

- **Métodos hierárquicos:** uma vez incluída num *cluster*, essa observação já não poderá sair do mesmo e, além disso, o número de grupos não é conhecido no momento inicial.

  Estes podem ainda ser classificados em dois tipos:

  - **Métodos aglomerativos:** são juntos grupos a cada iteração até existir apenas um *cluster* com todas as observações;
  - **Métodos divisivos:** o algoritmo inicia-se com todas as observações num único *cluster*, sendo estas separadas a cada iteração do procedimento;

- **Métodos não hierárquicos:** é definida uma divisão inicial das observações, sendo que estas podem mudar de grupo em qualquer momento no decorrer do algoritmo.

Uma boa análise de *clusters* é aquela que cria grupos com observações muito semelhantes entre si mas pouco semelhantes às observações de outros grupos. Para iniciar este processo, será então necessário clarificar o que são observações semelhantes ou distintas.

Para isso, é crucial começar por construir uma matriz de distâncias ou de dissemelhanças uma vez que esta irá definir como é calculada a similaridade entre duas observações.

Existem vários métodos para definir esta medida de distância, neste caso será estudado o método das distâncias Euclidianas entre dois pontos x e y (d), dado por:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{2.1}$$

Esta função tem as seguintes propriedades:

- $d(x,y) \geq 0, \forall(x,y)$
- $d(x,y) = 0$, se e só se $x = y$;
- $d(x,y) = d(y,x)$, $\forall(x,y)$;
- $d(x,y) \leq d(x,z) + d(z,y)$, $\forall(x,y,z)$;
- $d(x,y) = \max(d(x,z), d(y,z))$, $\forall(x,y,z)$.

2.1 Método de *k-means*

Neste estudo será utilizado o método de *k-means* (MacQueen et al., 1967). Este é classificado como um método não hierárquico, uma vez que consiste numa separação inicial das observações por um número de grupos escolhido previamente. Este número poderá ser definido através de um estudo antecipado ou simplesmente ao acaso.

Cada *cluster* é representado pelo seu centróide, ou seja, pela média de todas as observações pertencentes a esse *cluster*. O objetivo principal deste método consiste em minimizar a soma de todas as distâncias entre cada elemento e o respetivo centróide. O algoritmo só termina quando já não houverem alterações significativas nestas distâncias.

Além disso, os seus elementos podem mudar de grupo ao longo da implementação do algoritmo com vista a melhorá-lo, diminuindo assim a variabilidade.

A variabilidade total pode ser decomposta em dois tipos:

- **Variabilidade dentro dos grupos**
- **Variabilidade entre grupos**

Como outro qualquer, este método apresenta vantagens e desvantagens. A sua maior vantagem é a capacidade de aplicação a amostras de grande dimensão. No entanto, não devolve *clusters* do mesmo tamanho.

O algoritmo de *k-means* pode então ser resumido em 5 passos (DataNovia, s.d.):

1. Escolha do número de *clusters* (k) que deverão ser criados;
2. Seleção aleatória de k observações da amostra que serão os centróides dos *clusters* iniciais;

3. Associar cada observação ao centróide mais próximo, com base na distância Euclidiana entre a observação e o centróide;

4. Para cada um dos k *clusters*, atualizar o seu centróide através do cálculo da nova médias de todas as observações desse *cluster*;

5. Repetir os passos 3 e 4 até que não hajam mais alterações entre *clusters*, ou seja, até que se atinja a convergência.

Para desenvolver este algoritmo, será então utilizado o *software* estatístico *Rstudio* através dos passos seguintes:

1. Normalização dos dados de maneira a que as variáveis sejam comparáveis. Este processo consiste em transformar as mesmas de modo a que tenham média nula e desvio-padrão um. Para tal, foi utilizada a função *scale*;

2. Estimação do número ótimo de *clusters* a considerar através da função *fviz-nbclust*. Existem várias formas de determinar esse número, neste caso será utilizado o chamado método de *Elbow*:

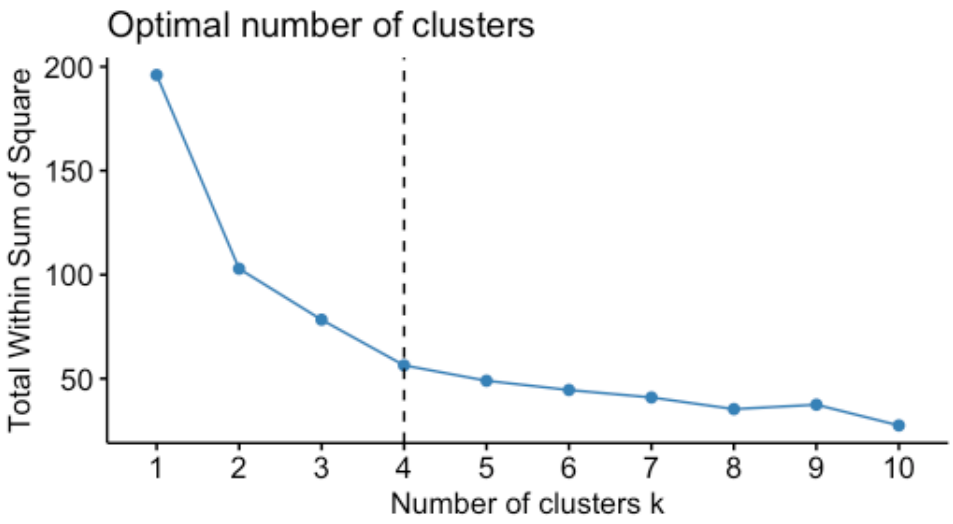

Figura 2.1: Exemplo do número ótimo de *clusters*

Através da figura 2.1 pode concluir-se que, à medida que o número de *clusters* aumenta, diminui a variabilidade entre os mesmos e mais do que 4 *clusters* não trarão melhorias ao algoritmo. Assim, deverá seguir-se a implementação do algoritmo com k = 4 no caso do exemplo apresentado;

3. Determinação dos *clusters* com base no método de k-means;

4. Vizualização dos pontos agrupados por *clusters*.

2.2 Análise de resultados

No presente capítulo serão apresentados os resultados obtidos através da aplicação do método de *k-means*, descrito em 2.1.

O algoritmo foi aplicado ao consumo médio mensal, idade e região de uma amostra aleatória de 5.000 clientes com o objetivo de averiguar se existe alguma relação entre essas características e a vinculação do cliente à plataforma digital.

Uma vez que a amostra é bastante extensa, os dados de consumo assim como a idade serão agrupados de modo a tornar a análise mais simples.

O número de *clusters* será escolhido conforme definido anteriormente:

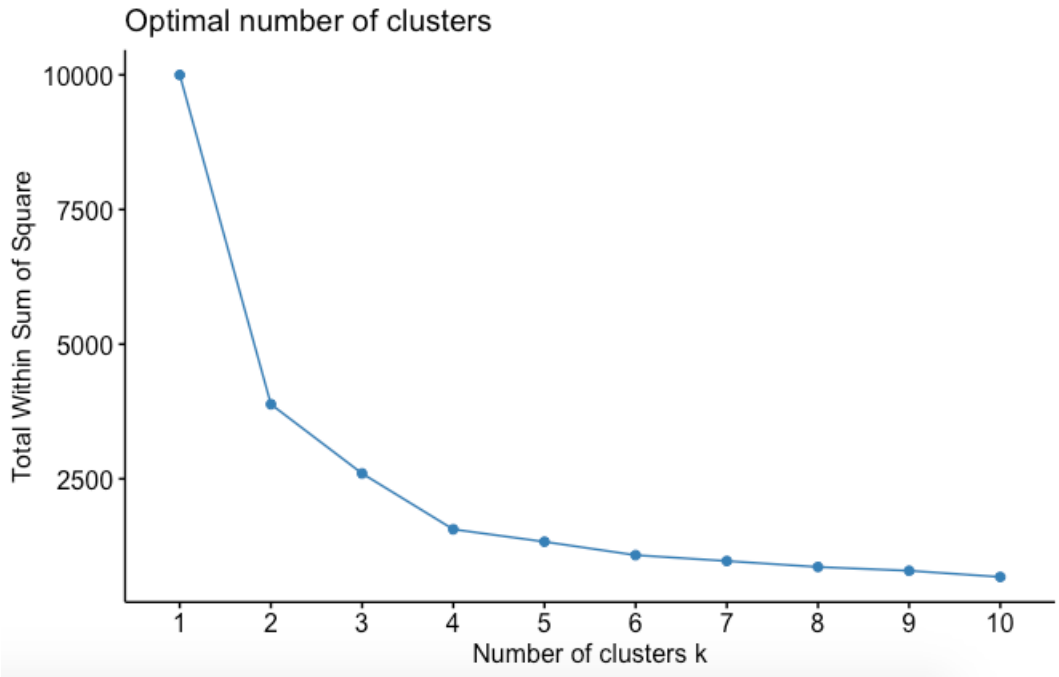

Figura 2.2: Número ótimo de *clusters*

Através da figura 2.2 verifica-se que, a partir de k=4, não há melhorias significativas no que diz respeito à variabilidade dentro dos grupos. Assim, o algoritmo será iniciado com 4 *clusters*.

A aplicação deste método permitiu então a formação dos *clusters* seguintes:

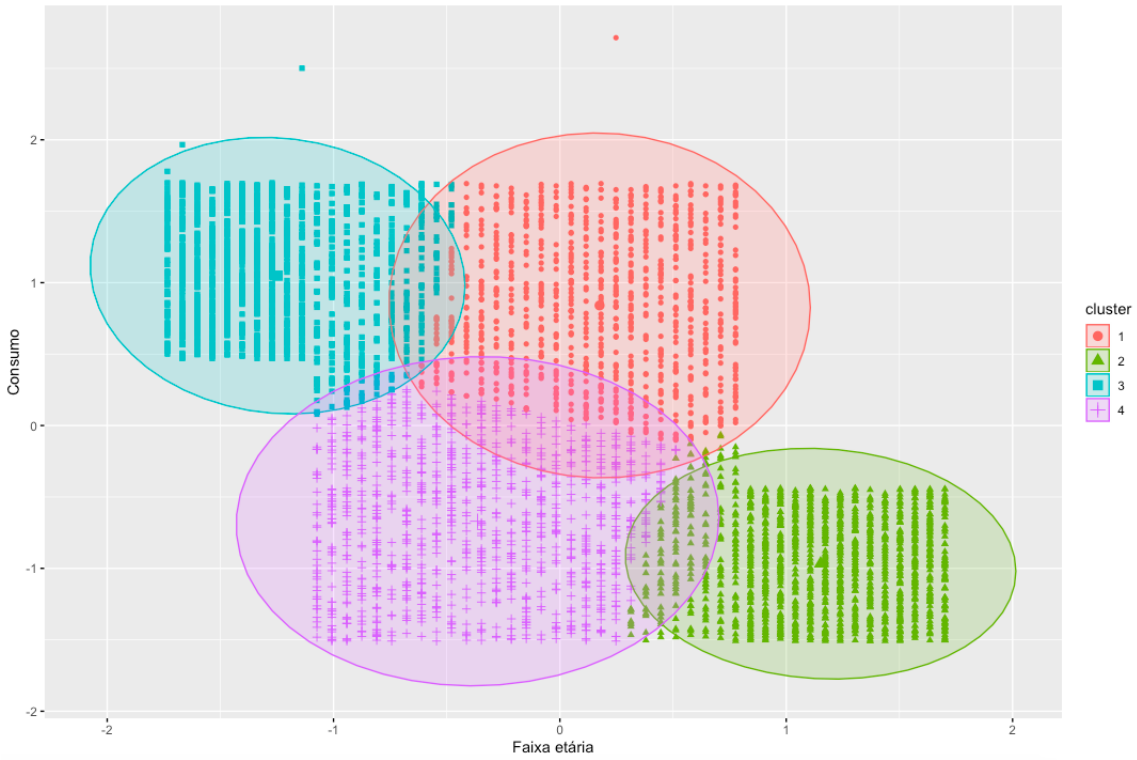

Figura 2.3: Número ótimo de *clusters*

Os *clusters* formados têm as seguintes dimensões:

- ***Cluster* 1:** 1048 elementos
- ***Cluster* 2:** 1586 elementos
- ***Cluster* 3:** 1302 elementos
- ***Cluster* 4:** 1064 elementos

Tem-se então que os *clusters* 2 e 3 são os de maior dimensão. No *cluster* 2 encontram-se os clientes de uma faixa etária superior com consumos mais reduzidos enquanto que no *cluster* 3 estão clientes mais jovens mas com consumos superiores.

De um modo geral, é expectável que este cenário ocorra na medida em que indivíduos de uma faixa etária inferior tendem a ser mais descuidados com o consumo de energia enquanto que, a partir de certa idade mais avançada, as pessoas têm um consumo mais consciente assim como também deixam de utilizar com tanta frequência algumas das fontes mais gastadoras de energia.

Na tabela seguinte encontra-se o tipo de consumo por *cluster*:

Consumo / *Cluster*	1	2	3	4
Muito Alto	306	0	548	0
Muito Baixo	0	416	0	160
Baixo	0	754	0	334
Médio	219	416	48	570
Alto	523	0	706	0

Tabela 2.1: Tipo de consumo por *cluster*

Pela tabela 2.1 é então possível confirmar que é no *cluster* 3 que estão presentes os clientes com consumo médio mensal mais alto enquanto que no *cluster* 2 se encontra os clientes com consumo médio mensal mais baixo.

Passando agora à análise por idade, categorizou-se na seguinte forma:

Idade / *Cluster*	1	2	3	4
Até 30	0	0	899	0
Entre 30 e 45	275	0	403	718
Entre 45 e 60	773	367	0	346
Superior a 60	0	1219	0	0

Tabela 2.2: Idade por *cluster*

Analisando a tabela 2.2 verifica-se que, confirmando o que se referiu anteriormente, é no *cluster* 2 que estão os clientes de uma faixa etária superior, enquanto que no *cluster* 3 ficam os indivíduos mais jovens.

Passando agora à análise por região, encontra-se na figura seguinte o volume de clientes por região e por *cluster*:

Região / *Cluster*	1	2	3	4
Estremadura	353	510	432	345
Alentejo	38	66	49	36
Beira	179	265	214	195
Algarve	76	110	84	74
Minho	366	573	489	382
Trás os Montes	36	62	34	32

Tabela 2.3: Região por *cluster*

Relativamente a esta característica do cliente, verifica-se na tabela 2.3 que não existem diferenças

muito significativas no que toca à região.

No entanto, a questão crucial é perceber onde se encontram afinal os clientes digitais.

Cluster	Nº de clientes digitais
1	272
2	180
3	1068
4	527

Tabela 2.4: Volume de clientes digitais por *cluster*

Olhando para a tabela 2.4, é bastante claro que mais de metade da proporção de clientes digitais se encontra no *cluster* 3.

Assim, conclui-se que o perfil do cliente digital da empresa assenta nos seguintes pontos:

- indivíduos jovens, principalmente com idade inferior a 30 anos;
- consumos mais altos, nomeadamente a partir dos 450 kW mensais.

Com os resultados obtidos é possível gerir de forma mais eficiente o tipo de campanhas de *marketing* enviadas. Isto porque, cada *cluster* permite conhecer melhor o cliente para o qual é enviada essa campanha.

Capítulo 3

Análise Exploratória dos Dados

Este capítulo será essencialmente composto por uma breve análise do ponto de situação da empresa assim como dos resultados obtidos no ano anterior de modo a encontrar determinados padrões no comportamento do utilizador da plataforma digital.

Para que isto seja possível, é fulcral começar por conhecer as operações disponíveis na mesma, ou seja, entender para que servem, se são muito ou pouco utilizadas e de que forma competem com as operações feitas através de outros canais alternativos.

3.1 Clientes

Durante largos anos, a maior parte das empresas não demonstrava qualquer tipo de preocupação com a definição de perfil de cliente por pensarem que a análise do público-alvo representaria apenas um elevado gasto de tempo e dinheiro. Assim, não existindo uma segmentação, admitia-se que se devia oferecer e publicitar os mesmos produtos a qualquer tipo de cliente. Com a evolução do mundo digital, a necessidade de conhecer o público-alvo da empresa começou a ser cada vez mais importante no sentido de orientar o desenvolvimento de produtos para cada cliente.

Assim sendo, o papel do cliente tem-se tornado fundamental para o crescimento da empresa na medida em que são estes que ditam o sucesso ou fracasso da mesma uma vez que têm o poder de compra do seu lado. Por mais que se criem novos produtos e tecnologias, é certo que se não existirem clientes não existe negócio nem criação de valor. É, por esse motivo, cada vez mais importante que a empresa se coloque no lugar destes e possa dessa forma compreender quais são as suas necessidades de modo a agir em concordância com as mesmas. Torna-se então imprescindível que a empresa construa uma relação forte de confiança com o cliente de maneira a aumentar o seu grau de satisfação e a conseguir manter uma relação contratual com os mesmos.

Com o passar dos anos e com a evolução tecnológica, o cliente espera cada vez mais um serviço rápido e até quase instantâneo. Consequentemente, a qualidade desse serviço tem vindo a ser mais valorizada relativamente ao preço, sendo o mau atendimento por parte da empresa uma das maiores causas de perda de clientes.

A vinculação de um cliente à empresa pode ser algo bastante frágil e é por isso importante não esquecer que um cliente insatisfeito poderá decidir abandonar a empresa além de que irá inevitavelmente partilhar a sua má experiência com outros potenciais clientes, reduzindo assim a probabilidade de

também estes adquirirem algum serviço junto da empresa.

Hoje em dia, também as redes sociais têm um papel cada vez mais importante nas escolhas dos consumidores na medida em que é muito comum que estes partilhem as suas experiências através das mesmas. É, assim, muito mais provável que o *feedback* se espalhe rapidamente, seja ele positivo ou negativo.

Desta forma, conhecer o cliente permite adaptar a oferta da empresa às necessidades de cada um assim como alinhar estas necessidades com os objetivos do seu negócio. Assim, faz todo o sentido começar por fazer uma análise do comportamento do cliente, da empresa e da sua interação com a mesma ao longo do período em estudo.

3.1.1 Ponto de situação de clientes em carteira a dezembro 2018

É evidente que nem toda a carteira de clientes utiliza a plataforma digital, optando assim por utilizar os canais alternativos, como lojas, agentes ou *contact center*.

Numa primeira abordagem, a empresa optou por fazer uma divisão da carteira de clientes com vista a segmentá-la pelos seus diferentes grupos, sendo estes:

- **Clientes totais**: clientes na carteira
- **Utilizadores registados**: clientes registados na plataforma digital
- **Utilizadores ativos**: clientes ativos na plataforma digital
- **Com login nos últimos 3 meses**: clientes ativos na plataforma digital que tenham acedido à mesma nos últimos 3 meses.

Apenas os utilizadores ativos conseguirão utilizar a plataforma digital de modo a realizar as operações que necessitarem, uma vez que nem todos os clientes que se registam chegam a ativar o seu registo e por sua vez não irão conseguir aceder à plataforma.

Esta divisão permite analisar de forma mais fácil a carteira assim como perceber se a plataforma está a ser utilizada com frequência no curto prazo e apresenta-se da seguinte forma:

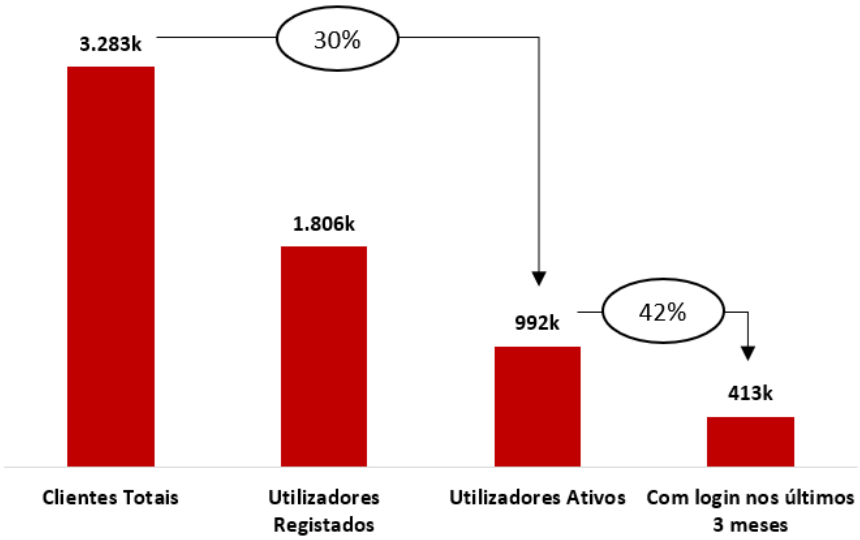

Figura 3.1: Clientes na carteira a 31 de dezembro de 2018

Como se pode verificar na figura 3.1, é bastante notória a diferença entre os clientes totais na carteira da empresa e os utilizadores que estão efetivamente ativos na plataforma, ou seja, os clientes digitais.

Desta forma conclui-se que, apesar da evolução no mundo do digital, existe ainda alguma resistência a este tipo de tecnologias por parte de alguns clientes, uma vez que apenas 30% destes são considerados digitais. Isto porque muitas pessoas mostram ainda ter alguma falta de confiança no digital para resolver temas críticos.

Note-se que, existem mais clientes registados do que ativos uma vez que muitos deles não terminam o processo de ativação, ou seja, fazem o registo mas acabam por não seguir para a ativação do mesmo. Pode ainda concluir-se que, deste grupo de clientes digitais, só cerca de 42% acedeu nos últimos 3 meses à sua área de cliente.

Com este breve resumo, pode verificar-se que é ainda muito significativa a quantidade de clientes que não utiliza a plataforma, seja por não sentir necessidade ou até mesmo por mero desconhecimento da existência da mesma. Além disso, o facto de menos de metade dos utilizadores ativos terem efetuado, nos últimos 3 meses, qualquer interação com a empresa através da plataforma digital leva a crer que, apesar de se terem registado, estes clientes não sentem necessidade de a utilizar ou até mesmo que não sabem o que e como o podem fazer.

3.1.2 Visão anual de clientes

Importa agora perceber qual foi o comportamento dos clientes ao longo de todo o período em estudo. Para tal, será analisada a evolução do número total de clientes na carteira, dos utilizadores ativos e dos utilizadores que efetuaram *login* nos últimos 3 meses durante esse período, representada na figura seguinte:

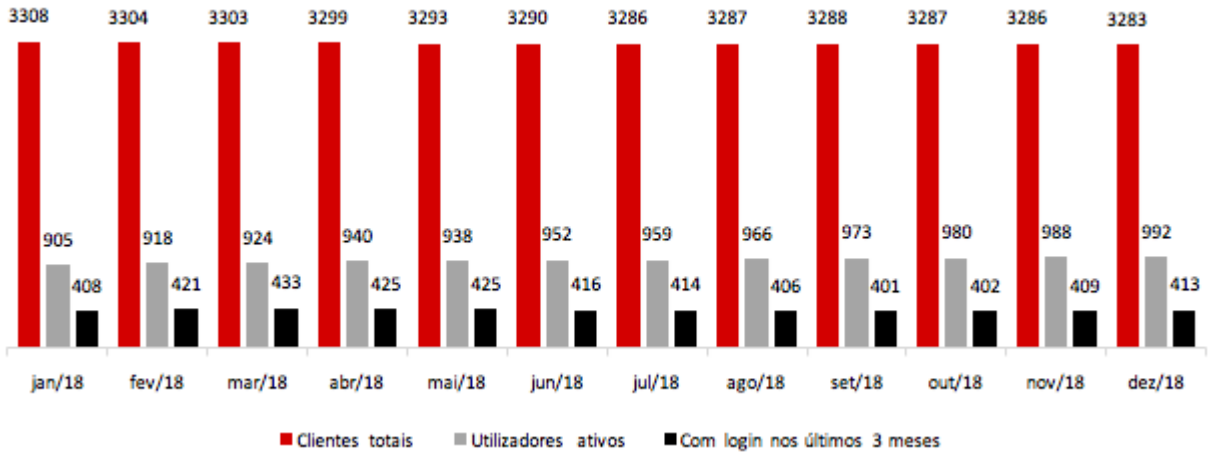

Figura 3.2: Visão geral de clientes (k)

Tal como se pode ver na figura 3.2, o número de clientes na carteira tem vindo a diminuir constantemente ao longo do período em estudo, ainda que de forma pouco significativa, sendo este decréscimo de apenas 0,07% ao mês.

Contrariamente a esta variação e ao que seria de esperar, é de notar que o número de utilizadores ativos no mesmo período de tempo aumentou cerca de 0,8%. por mês

No geral, a evolução total não deixa de ser positiva pois significa que apesar de existirem menos clientes, essa perda tem sido compensada e até superada por aqueles que se tornaram digitais.

3.2 Contratos

Agora que ficou clara a importância do papel do cliente para a empresa, resta compreender a ligação entre os dois, ou seja, a sua relação contratual. O cliente formaliza esta relação com a empresa sempre e quando contrata um produto e/ou serviço à mesma.

Uma vez que a empresa dispõe de vários serviços distintos, os clientes podem também adquirir mais do que um contrato, ou até mesmo gerir contratos de outros.

O cliente pode então contratar os seguintes serviços:

- Eletricidade;
- Eletricidade Verde;
- Gás;
- Gás Natural;
- Mobilidade Elétrica.

3.2.1 Ponto de situação de contratos em carteira a dezembro 2018

Tal como os clientes da empresa, também os contratos podem ou não estar registados na plataforma digital.

A empresa optou por fazer uma segmentação semelhante à de clientes, sendo:

- **Contratos Ativos** – contratos ativos na carteira, ou seja, contratos em vigor até à data
- **Contratos Registados na Plataforma Digital** – contratos ativos e registados na plataforma digital, ou seja, contratos que estão a ser geridos na plataforma digital
- **Contratos com FE** – contratos ativos que aderiram à fatura eletrónica

Na figura seguinte encontra-se uma representação dessa segmentação:

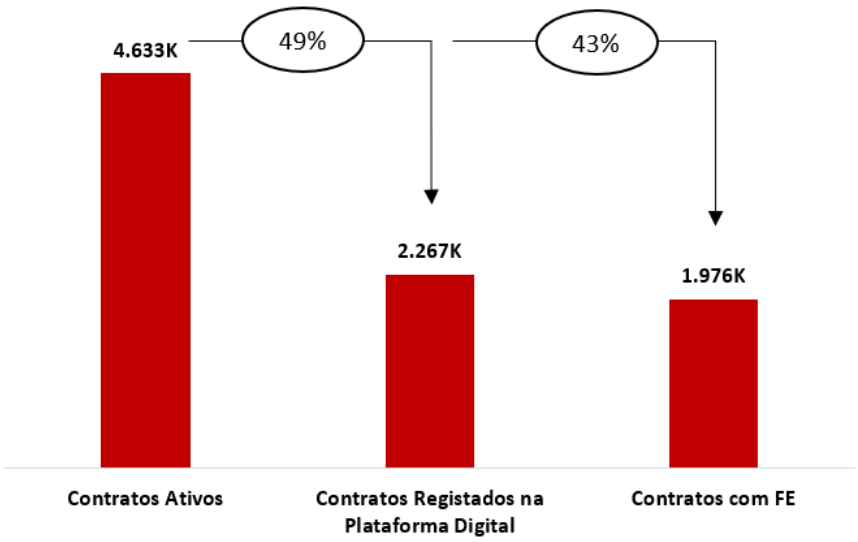

Figura 3.3: Contratos na carteira a 31 de dezembro de 2018

Analisando a figura 3.3 verifica-se que mais de metade dos contratos ativos na carteira não estão registados na plataforma digital.

De certa forma, esta diferença já seria expectável tendo em conta que 70% dos clientes da empresa não são digitais e, consequentemente, também não lhes será possível gerir os seus contratos na plataforma.

Por fim, observa-se que apenas 43% dos contratos ativos usufruem de fatura electrónica, cuja análise irá ser estudada mais à frente.

3.2.2 Visão anual de contratos

Resta assim analisar o comportamento dos contratos ativos e registados na plataforma ao longo de todo o ano.

Uma vez que o número de clientes totais na carteira tem vindo a diminuir, conforme foi possível verificar na figura 3.2, seria então também de esperar que o mesmo acontecesse com o número de contratos em vigor.

Será então analisada a evolução mensal do número total de contratos ativos e dos contratos que estão registados na plataforma durante esse período, representada na figura seguinte:

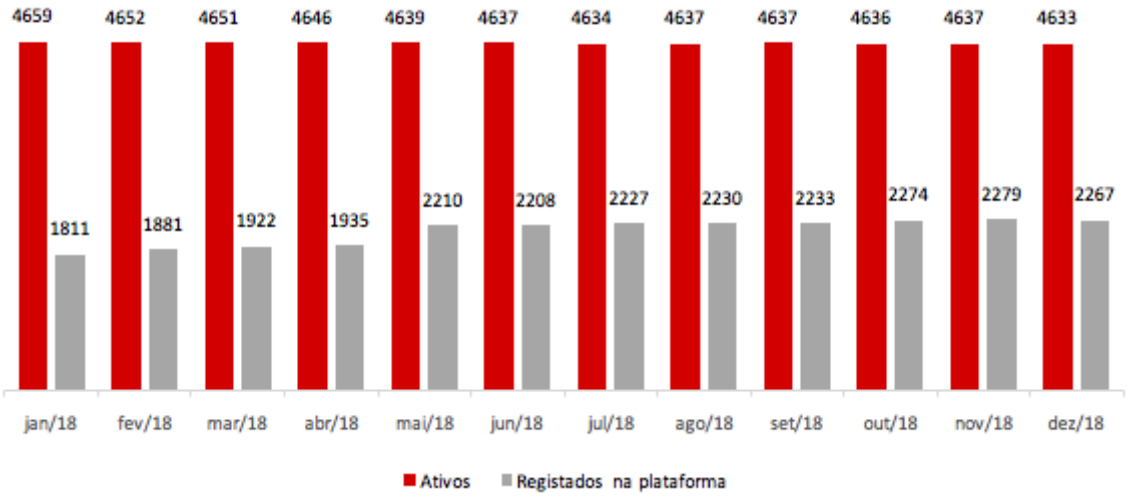

Figura 3.4: Visão geral de contratos (k)

Tal como já se previa, na figura 3.4 podemos ver que também o número de contratos ativos diminuiu cerca de 0,05% ao longo do ano de 2018.

No entanto, têm sido cada vez mais os contratos registados na plataforma digital. Durante o período em análise, o número de contratos geridos na plataforma digital aumentou 2% ao mês.

Concluindo, mais uma vez a diminuição dos contratos ativos foi compensada pelo ganho de contratos geridos na plataforma digital.

3.3 Operações

Como já havia sido referido anteriormente, são várias as operações disponíveis para o cliente na plataforma digital em estudo. Estas foram desenvolvidas e pensadas com base naquilo que seria útil para os seus utilizadores.

3.3.1 *Downloads* da aplicação

A plataforma digital pode ser utilizada através do *browser*, mas também através da aplicação da empresa.

Desta forma, permite-se ao utilizador usufruir de uma experiência bastante mais optimizada e mais apelativa para gerir a sua energia, uma vez que poderá aceder à mesma através de qualquer dispositivo móvel.

Através da aplicação poder-se-á, entre outras coisas:

- alterar detalhes do contrato;
- enviar leituras e receber notificações;
- consultar faturas e ver o valor que se tem a pagar;
- consultar histórico de consumos;
- pedir nova referência multibanco para pagamento ou enviar os dados de pagamento por SMS;
- efetuar download de faturas;
- efetuar novos contratos;
- aderir à fatura eletrónica e débito direto;
- mudar o plano de energia;
- efetuar pedidos de informação e reclamações;
- acompanhar os seus consumos e comparar com consumos anteriores ou com clientes com contratos semelhantes;
- pedir informações e consultar as perguntas mais frequentes na área de Apoio ao Cliente.

Na figura seguinte será então representada a evolução do número de *downloads* realizados ao longo do ano:

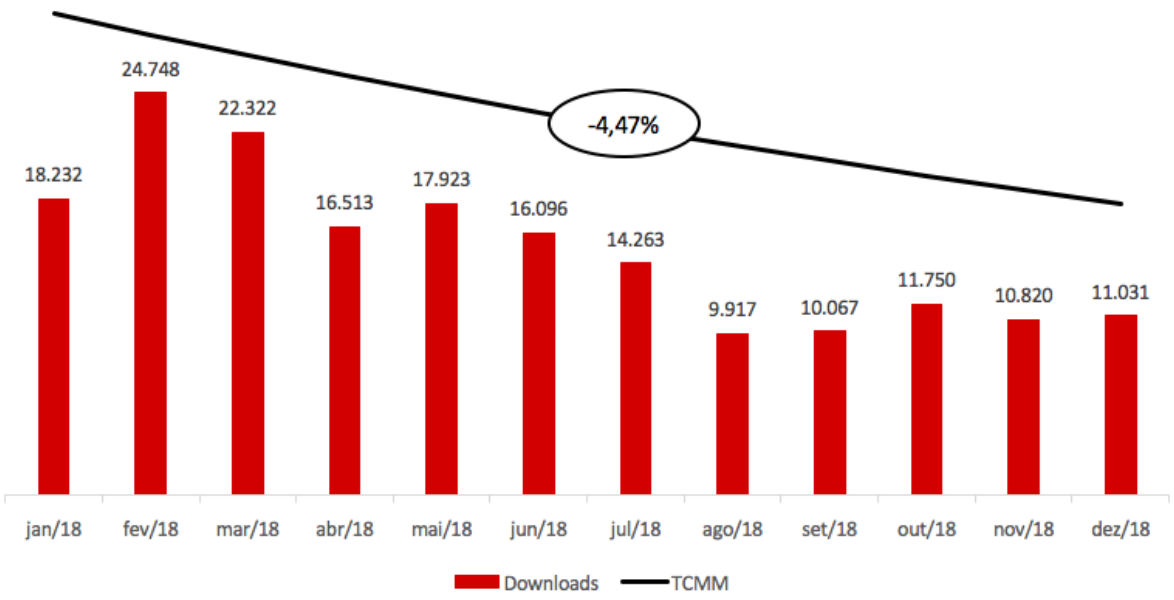

Figura 3.5: Número de downloads da aplicação (unidades)

Note-se que, de acordo com a figura 3.5, fevereiro e março foram os meses com maior volume de *downloads* enquanto que agosto e setembro foram meses mais fracos.

A Taxa de Crescimento Média Mensal (TCMM) é negativa, pelo que se pode concluir que, em média, o volume de downloads tem vindo a diminuir ao longo do último ano.

3.3.2 *Logins*

Após a ativação do registo, o utilizador poderá começar a utilizar a plataforma e realizar todas as operações que necessitar. É contabilizado um *login* sempre e quando o utilizador aceda à plataforma digital.

O número de *logins* efetuados ao longo do período em estudo encontra-se representado na figura seguinte:

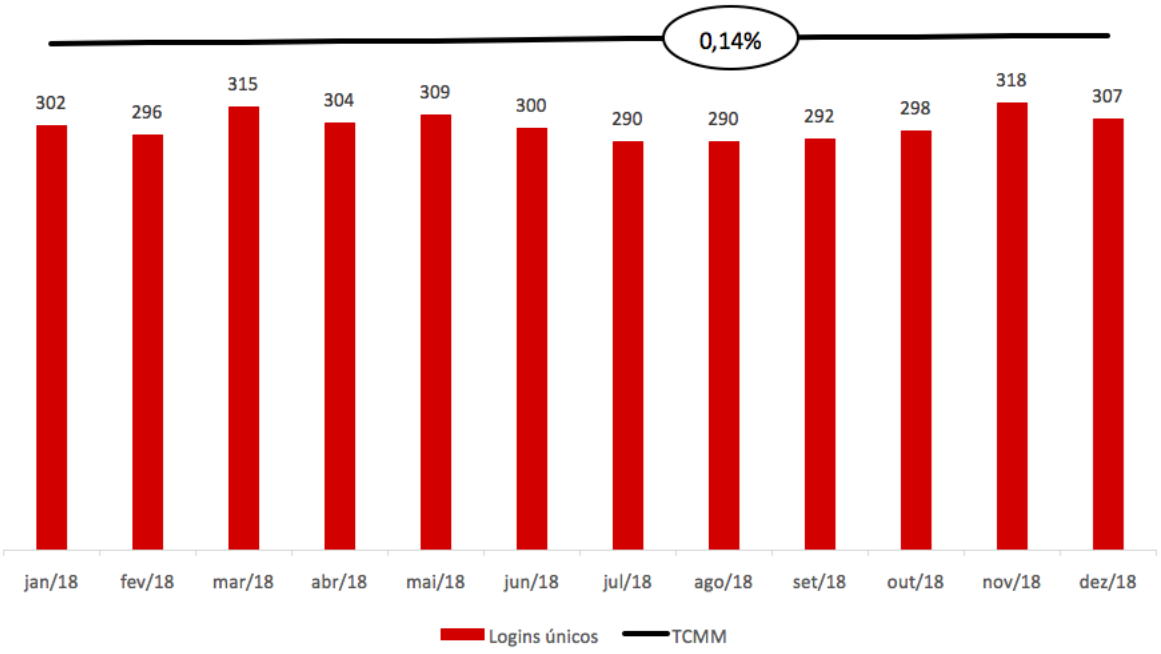

Figura 3.6: Número de logins (k)

Como se pode verificar na figura 3.6, o volume de *logins* oscilou bastante ao longo dos meses, no entanto, nota-se uma tendência de redução durante os meses de verão. Esta tendência é expectável na medida em que é uma altura do ano em que os consumidores costumam usufruir do seu período de férias e por isso estarão menos ativos na plataforma.

Avaliando pela TCMM, houve um aumento pouco significativo de 0,14%.

3.3.3 Comunicação de leituras

O operador da rede de distribuição tem a obrigação de assegurar a leitura do contador, no entanto, quando o distribuidor não consegue aceder ao contador, não consegue comunicar a leitura ao comercializador de energia. Assim, não dispondo de uma leitura real do contador na data de emissão de fatura, o comercializador será obrigado a recorrer a uma estimativa do consumo para poder dessa forma faturar o consumo.

Assim, a leitura do contador de eletricidade e/ou de gás natural deve ser enviada regularmente pelo cliente para garantir que o valor das suas faturas corresponde à energia que foi efetivamente consumida por si.

A evolução do volume de comunicações de leituras está representado na figura seguinte:

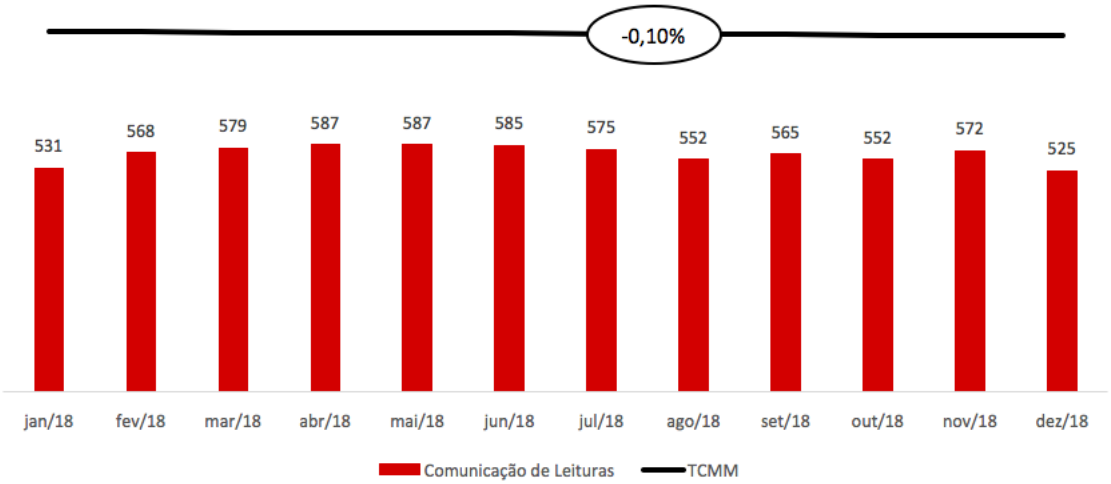

Figura 3.7: Número de comunicação de leituras (k)

Por observação da figura 3.7 pode concluir-se que não houve uma grande variação do volume de comunicação de leituras ao longo do ano, no entanto, verifica-se facilmente que janeiro e dezembro foram os meses em que menos comunicações se fizeram.

A TCMM foi de -0,10%, um decréscimo pouco significativo.

3.3.4 Consultas

Tal como já foi referido anteriormente, com o decorrer do tempo foram existindo cada vez mais funcionalidades na plataforma digital.

Desta forma, é possível efectuar vários tipos de consultas *online*, como seja:

- Consulta de faturas;
- Consulta de consumos;
- Consulta de leituras;
- Consulta de detalhes dos contratos;
- Consulta de documentos;
- Consulta de movimentos;
- Consulta de potências.

As consultas serão analisadas como um todo e a sua evolução mensal encontra-se representada na figura seguinte:

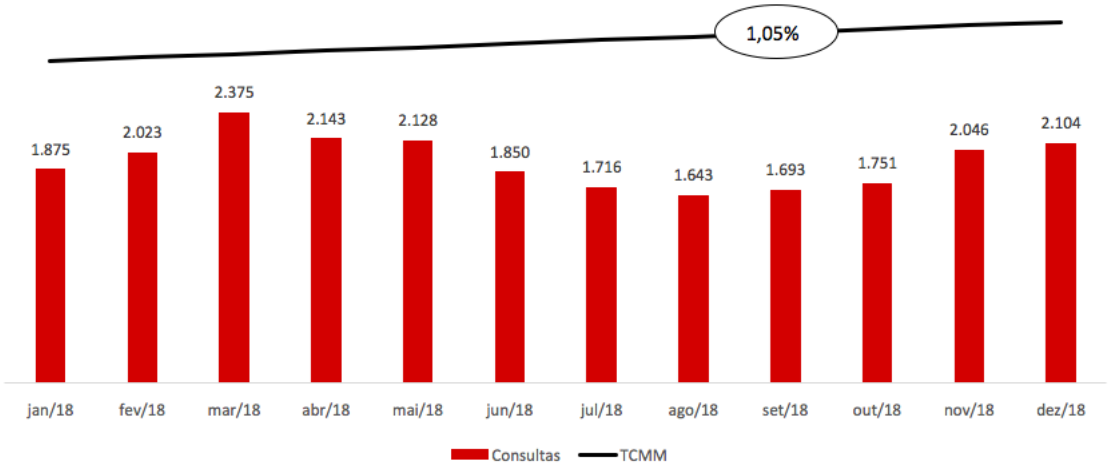

Figura 3.8: Número de consultas (k)

Analisando a figura 3.8, mais uma vez é bastante claro que nos meses de verão são efetuadas menos consultas por parte dos clientes.

Pela TCMM pode verificar-se que houve uma variação positiva de 1,05% ao longo dos meses de 2018.

3.3.5 Pedidos de Informação

Sempre e quando o consumidor queira pedir qualquer tipo de esclarecimento, tem a possibilidade de efetuar um pedido de informação para o efeito.

Na figura seguinte está representado o número de pedidos de informação ao longo do período em estudo:

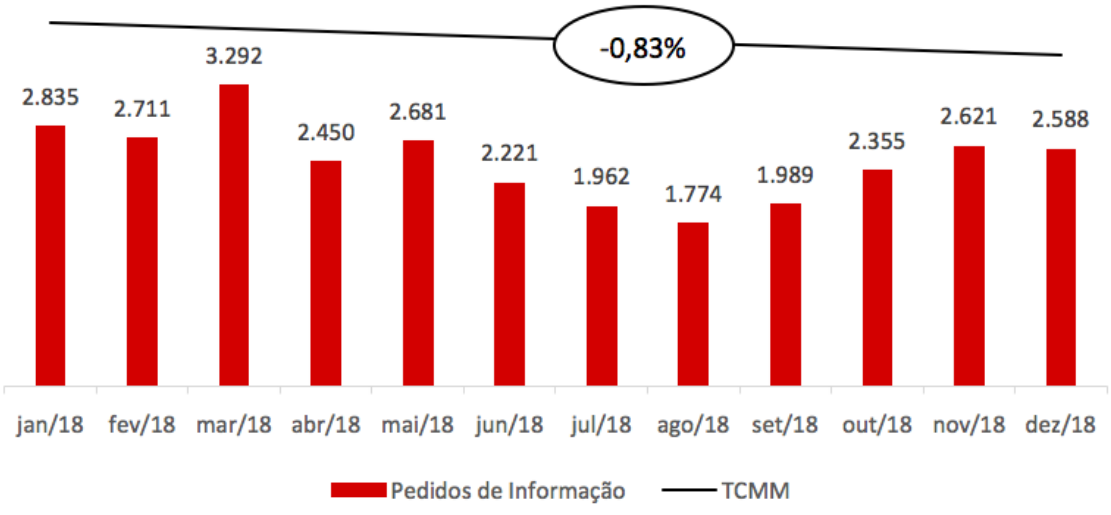

Figura 3.9: Número de pedidos de informação (k)

Através da figura 3.9, é possível verificar alguma tendência de decréscimo de pedidos de informação nos meses de Verão. Esta tendência poderá ser justificada pelo período de férias dos utilizadores, uma vez que passarão menos tempo nas suas habitações e menos probabilidade haverá de necessitarem de algum pedido de informação.

Segundo a TCMM, estes têm vindo a diminuir 0,83% ao longo do ano.

3.3.6 Reclamações

Para além dos pedidos de informação, existe também a possibilidade de efetuar uma reclamação quando o cliente não estiver satisfeito com alguma situação.

A evolução mensal do volume de reclamações efetuadas encontra-se representado na figura seguinte:

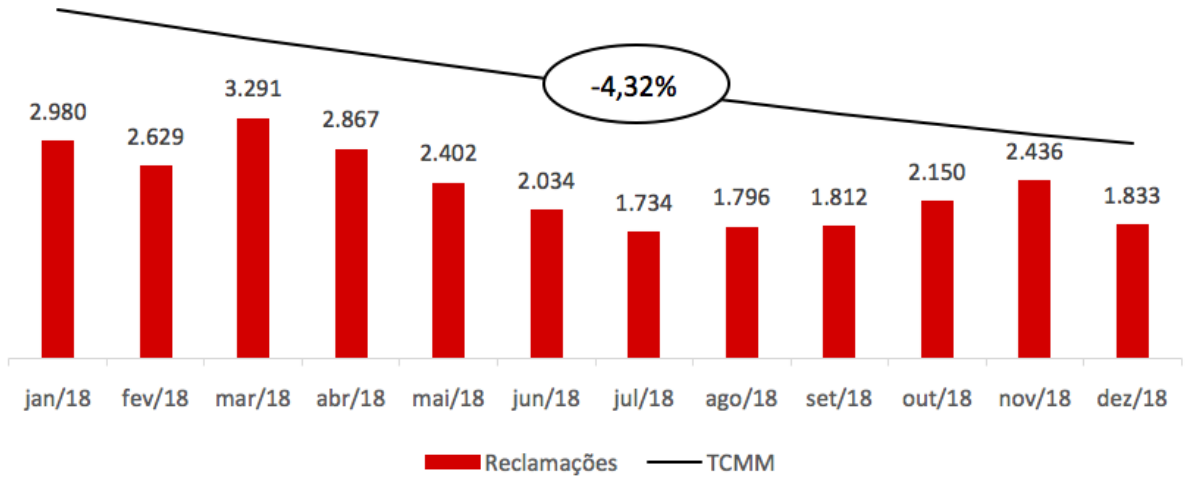

Figura 3.10: Número de reclamações

Conforme se observa na figura 3.10, também no volume de reclamações se nota uma diminuição significativa durante os meses de verão. Esta tendência poderá mais uma vez ser justificada pelo período de férias dos consumidores.

Segundo a TCMM, o número de reclamações tem vindo a diminuir 4,32% ao longo de todo o ano, o que é obviamente bastante favorável à empresa.

3.4 Digital *vs* Outros canais

Apesar da existência da plataforma digital, os clientes da empresa podem realizar operações noutros canais, como seja: lojas, agentes e *contact center*.

A verdade é que são diversas as vantagens de utilizar uma plataforma capaz de satisfazer as necessidades do cliente em qualquer hora e em qualquer lugar, sem filas ou limitações geográficas. No entanto,

há certas características e comodidades de uma loja que, para alguns clientes, nunca serão superadas por uma plataforma *online*.

Assim, é importante fazer uma quantificação do número de clientes ganhos aos outros canais alternativos, por operação realizada. Deste modo será possível perceber de que maneira a criação da plataforma veio impactar o modo de interação dos clientes com a empresa.

Para tal, torna-se necessário estabelecer quais as operações mais frequentes de modo a compreender se estas estão correlacionadas entre si.

3.4.1 Comunicação de Leituras

O envio de leituras pode ser feito de diversas formas, como seja: linha telefónica, plataforma digital, lojas e agentes. Quando a leitura é efectuada pelos outros canais, existe um risco maior desta não ser contabilizada a tempo da fatura seguinte, uma vez que o processo não é automático.

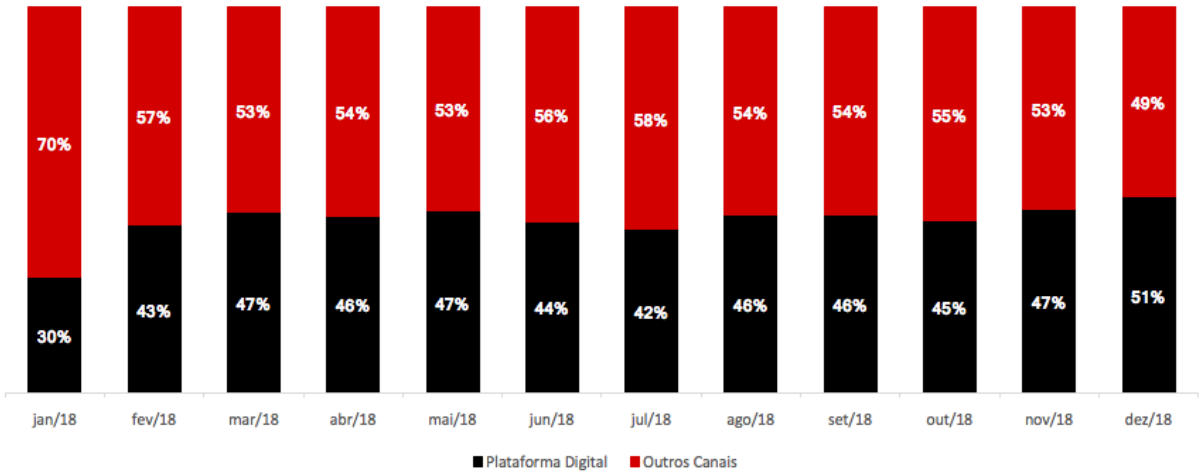

Figura 3.11: Comunicações de leitura na plataforma digital *vs* noutros canais

O número de leituras comunicadas através da plataforma digital tem vindo a aumentar. No mês de dezembro, o volume destas foi de 54% do total de leituras enviadas, enquanto que, no início do ano, tinha sido de apenas 30%.

3.4.2 Pedido de Referência Expresso

O pedido de referência expresso é efectuado sempre e quando o cliente necessitar de uma nova referência para pagamento no multibanco, caso tenha deixado passar a data limite de pagamento da sua fatura, do aviso de interrupção ou de outro aviso de dívida.

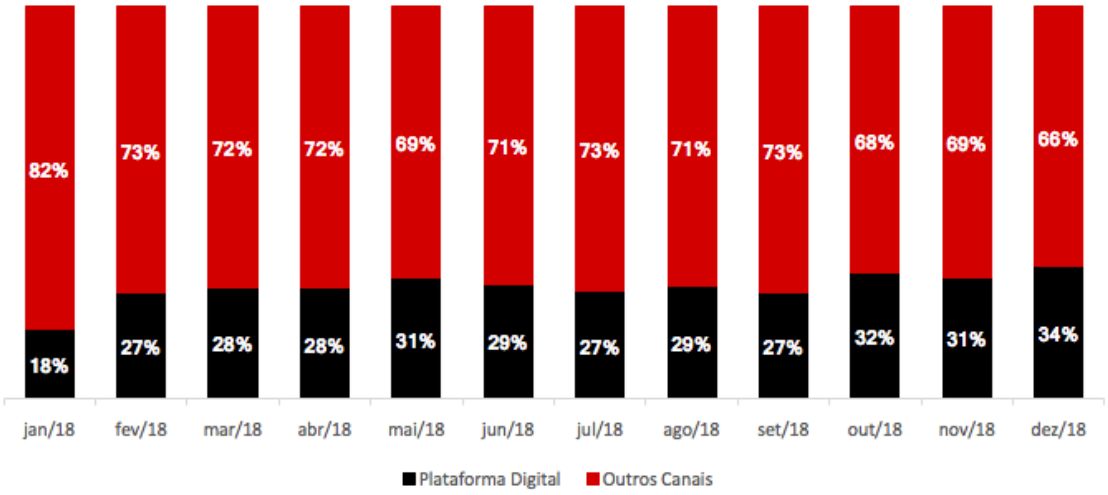

Figura 3.12: Pedidos de referência expresso na plataforma digital *vs* noutros canais

Conforme se pode ver na figura 3.12, também o volume de pedidos de referência expresso através da plataforma digital aumentou no decorrer no período em estudo.

No início do ano apenas 18% destes pedidos teriam sido efectuados digitalmente, chegando este volume aos 34% em dezembro.

3.4.3 Adesão à Fatura Eletrónica

A fatura eletrónica é igual à fatura que chega por correio, cumprindo todas as exigências legais. A única coisa que difere é a forma de envio, sendo esta enviada por *e-mail*.

Por este motivo, torna-se bastante apelativo para a empresa que haja uma grande adesão a esta funcionalidade uma vez que irá certamente reduzir os seus custos com os envios de faturas para além de evitar um grande desperdício de papel.

A evolução das adesões à fatura eletrónica está representada na figura seguinte:

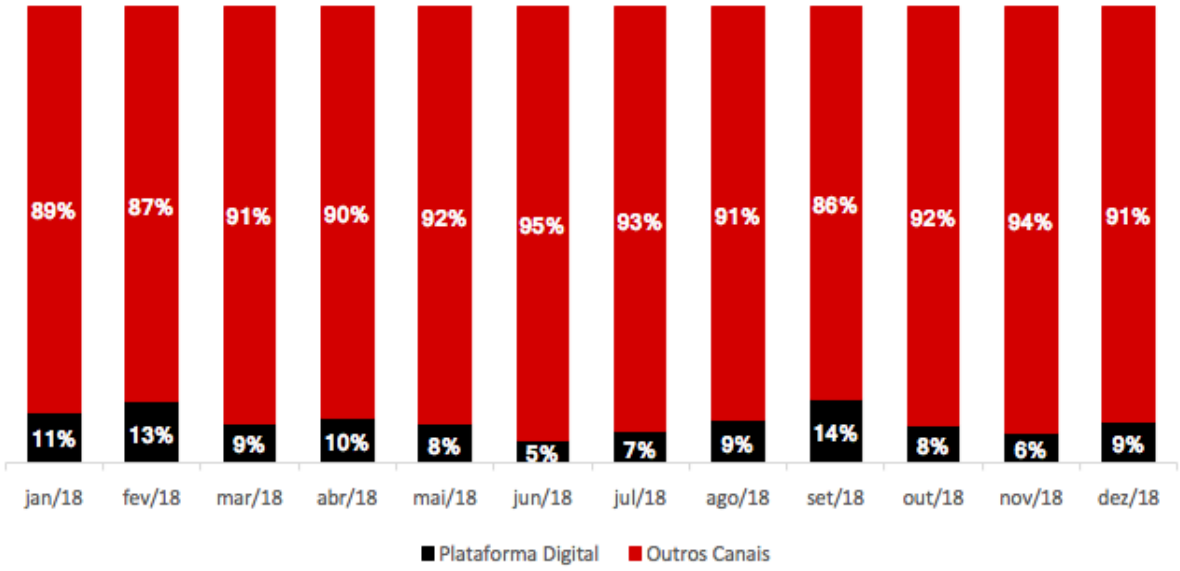

Figura 3.13: Adesões à fatura eletrónica na plataforma digital *vs* noutros canais

Por observação da figura 3.13, pode facilmente concluir-se que o volume de adesões à fatura electrónica é bastante inferior na plataforma digital relativamente a lojas, agentes ou *contact center*. Além disso, é ainda possível verificar que este volume tem vindo a diminuir ao longo do ano de 2018.

Importa referir que, não sendo possível fazer uma distinção entre as adesões no momento da contratação e as adesões posteriores à contratação, fica pouco claro se este volume depende ou não do momento em que o cliente adere à fatura electrónica.

Se assim for, isto explicaria o facto do volume de adesões nos outros canais ser bastante superior ao da plataforma digital uma vez que a contratação é, maioritariamente, feita em loja ou agentes.

3.4.4 Adesão ao Débito Direto

O débito direto é um serviço de pagamento que, tal como o seu nome indica, consiste em debitar diretamente da conta do cliente o valor devido, com a periodicidade escolhida.

Assim, o cliente garante o pagamento atempado das suas faturas sem preocupações nem perdas de tempo.

As adesões ao débito direto, por mês, encontram-se representadas na seguinte figura:

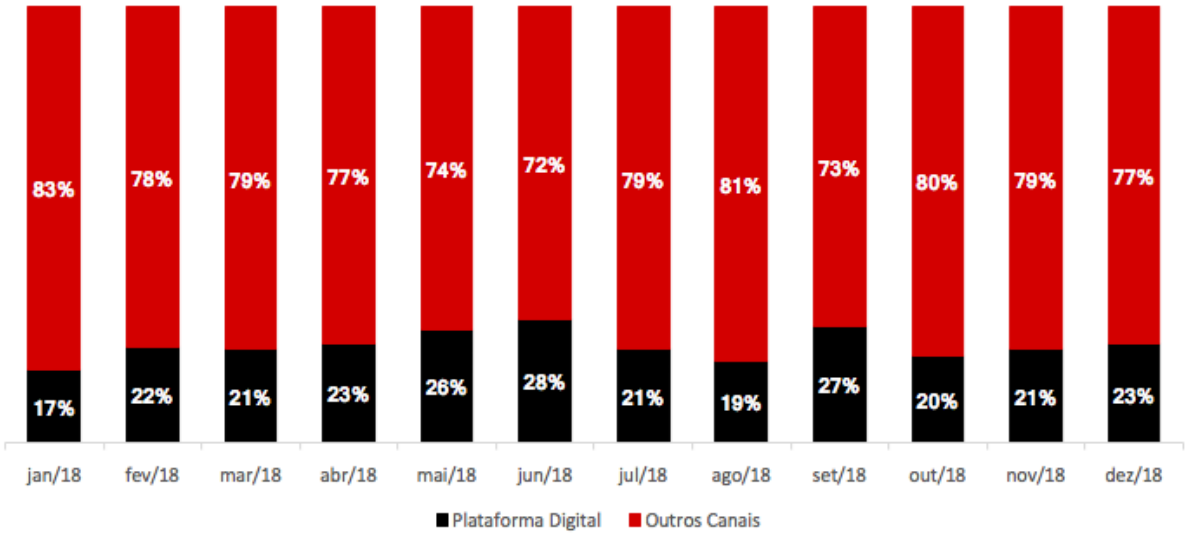

Figura 3.14: Adesões ao débito direto na plataforma digital *vs* noutros canais

Conforme se observa na figura 3.14, também o volume de adesões ao débito direto é significativamente menor na plataforma digital comparativamente a outros canais.

Mais uma vez, não é possível clarificar quando é que a adesão é efectuada, relativamente ao momento da contratação.

Supondo que o cliente adere ao débito direto no momento da contratação, é expectável que estes volumes não tenham uma grande oscilação.

Capítulo 4

Análise de Sazonalidade

É do conhecimento geral que um dos maiores fatores com impacto no *marketing* digital é o efeito de sazonalidade. Dependendo do tipo de negócio, é expectável que haja alturas do ano mais favoráveis à empresa do que outras.

Uma vez que se trata de uma empresa no setor da energia, é provável que para além de algumas operações disponíveis na plataforma, o consumo de energia seja significativamente diferente em determinados meses devido à alteração do clima ao longo do ano.

Assim, para verificar a existência ou não de sazonalidade, será feita uma análise da variância seguida de testes *pairwise*.

4.1 Análise da Variância Simples (*One Way* ANOVA)

A análise de variância (Alpuim, 2017) consiste em verificar se as médias de três ou mais grupos têm ou não diferenças significativas entre si.

No fundo, esta análise baseia-se numa generalização da estatística T, uma vez que existe um maior número de amostras.

Suponha-se que se têm amostras (X_{i1}, X_{i2}, ..., X_{in_i}), i = 1, ..., I, com dimensão n_i, independentes entre si e provenientes de populações normalmente distribuídas com médias μ_i e igual variância σ^2.

O objetivo será então testar as hipóteses:

H_0: $\mu_1 = \mu_2 = ...\mu_I$ vs H_1: existe pelo menos um μ_i que se distingue dos restantes

Assim, cada uma das observações pode escrever-se da seguinte maneira:

$$X_{ij} = \mu_i + \varepsilon_{ij}, i = 1, ..., I, j = 1, ...n_i$$

onde μ_i são constantes e ε_{ij} são variáveis aleatórias independentes e identicamente distribuídas e seguem uma distribuição normal de média 0 e variância σ^2.

Será doravante designado por N, o número total de observações, ou seja:

$$N = n_1 + n_2 + ... + n_{\mathrm{I}} = \sum_{i=1}^{I} n_{\mathrm{i}}$$

4.1.1 Partição da Soma de Quadrados

Seja $\bar{\mathrm{X}}_{\mathrm{i}}$ a média de todas as observações pertencentes à amostra i:

$$\overline{X}_{\mathrm{i}} = \frac{X_{\mathrm{i}1} + X_{\mathrm{i}2} + ... + X_{\mathrm{in_i}}}{n_{\mathrm{i}}} = \frac{1}{n_{\mathrm{i}}} \sum_{j=1}^{n_{\mathrm{i}}} X_{\mathrm{ij}}$$

Seja $\bar{\mathrm{X}}$ a média global de todas as observações das I amostras:

$$\overline{X} = \frac{n_1\overline{X}_1 + n_2\overline{X}_2 + ... + n_{\mathrm{I}}\overline{X}_{\mathrm{I}}}{N} = \frac{1}{N} \sum_{i=1}^{I} \sum_{j=1}^{n_{\mathrm{i}}} X_{\mathrm{ij}}$$

Generalizando para I populações, a variabilidade entre grupos pode ser escrita como:

$$SQ_{\mathrm{ext}} = \sum_{i=1}^{I} n_{\mathrm{i}}(\overline{X}_{\mathrm{i}} - \overline{X})^2$$

A variabilidade total da amostra é definida por:

$$SQ_{\mathrm{Tot}} = \sum_{i=1}^{I} \sum_{j=1}^{n_{\mathrm{i}}} (X_{\mathrm{ij}} - \overline{X})^2$$

A variabilidade dentro dos grupos é definida por:

$$\mathrm{SQ}_e = \sum_{i=1}^{I} \sum_{j=1}^{n_{\mathrm{i}}} (X_{\mathrm{ij}} - \overline{X}_{\mathrm{i}})^2$$

A variabilidade total da amostra tem-se como a soma entre a variabilidade dentro dos grupos e a variabilidade entre os grupos, ou seja:

$$\mathrm{SQ}_{Tot} = \mathrm{SQ}_e + \mathrm{SQ}_{ext}$$

Sob a validade da hipótese nula, a variável aleatória:

$$\frac{\mathrm{SQ}_{ext}/(I-1)}{\mathrm{SQ}_e/(N-I)}$$

tem distribuição F com I-1 e N-I graus de liberdade uma vez que se trata de um quociente entre dois qui-quadrados independentes sobre os respetivos graus de liberdade.

Assim, a estatística de teste é dada por:

$$F = \frac{\mathrm{SQ}_{ext}/(I-1)}{\mathrm{SQ}_e/(N-I)} = \frac{\mathrm{MQ}_{ext}}{\mathrm{MQ}_e}$$

que exprime a relação entre a variabilidade entre os grupos e a variabilidade dentro dos grupo.

A região de rejeição do teste será então:

$$F = \frac{\mathrm{MQ}_{ext}}{\mathrm{MQ}_e} > F^{1-\alpha}_{I-1,N-I}$$

que representa o quantil de ordem 1-α da distribuição F com I-1 e N-I graus de liberdade.

Para que os resultados do teste sejam de fácil compreensão, estes representam-se numa tabela com todos os cálculos auxiliares utilizados na construção da estatística de teste.

Esta tabela designa-se por ANOVA (ANalysis Of VAriance) e apresenta-se da seguinte maneira:

FONTE DE VARIAÇÃO	SOMA DE QUADRADOS	GRAUS DE LIBERDADE	MÉDIA DE QUADRADOS	ESTATÍSTICA *F*
Entre grupos	$SQ_{ext} = \sum_{i=1}^{I} n_i \left(\bar{X}_i - \bar{X}\right)^2$	*I*-1	$MQ_{ext} = \frac{SQ_{ext}}{I-1}$	$F=MQ_{ext}/MQ_e$
Erro	$SQ_e = \sum_{i=1}^{I}\sum_{j=1}^{n_i} \left(X_{ij} - \bar{X}_i\right)^2$	*N-I*	$MQ_e = \frac{SQ_e}{N-I}$	
Total	$SQ_{Tot} = \sum_{i=1}^{I}\sum_{j=1}^{n_i} \left(X_{ij} - \bar{X}\right)^2$	*N*-1	$MQ_{Tot} = \frac{SQ_{Tot}}{N-1}$	

Tabela 4.1: ANOVA

4.2 Testes *pairwise*

Apesar da análise de variância nos dar indicação de que pelo menos um grupo difere dos restantes no caso de se rejeitar H0, esta não nos permite saber qual deles é diferente.

Uma vez que se pretende identificar quais são exatamente os grupos que são diferentes dos outros, a ANOVA será seguida de testes para a comparação específica dos grupos, ou seja, dos chamados testes *pairwise*.

Obviamente que só fará sentido seguir para estes testes caso a ANOVA tenha revelado a existência de diferenças significativas.

É então fundamental encontrar todos os pares a comparar e, para isso, o número total de testes a efetuar é dado pela seguinte equação:

$$c = \frac{I \times (I-1)}{2}$$

Uma vez que a amostra em estudo é composta por 4 grupos, serão então efetuados 6 testes, conforme mostra a figura seguinte:

	T1	T2	T3	T4
T1				
T2	T2 vs T1			
T3	T3 vs T1	T3 vs T2		
T4	T4 vs T1	T4 vs T2	T4 vs T3	

Tabela 4.2: Pares a comparar

4.2.1 *Fisher's Least Significant Difference* (LSD)

Um dos métodos de comparação de pares utilizado será o chamado *Least Significant Difference* (Williams et al., 2010), desenvolvido por Fisher em 1935.

Este método consiste em utilizar a estatística T, onde:

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\mathrm{MQ}_e \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

e segue, sob a validade da hipótese nula, uma distribuição *t-de-student* com N-I graus de liberdade.

Para um determinado nível de significância alfa, t será considerada significativa caso o seu valor seja superior ao quantil encontrado através da distribuição t para o valor de alfa considerado, ou seja, se:

$$T > t_{\text{N-I},\alpha}$$

Assim, desenvolvendo as equações anteriormente mencionadas, podemos dizer que a diferença entre as médias é considerada significativa se:

$$|\overline{X}_1 - \overline{X}_2| > \text{LSD} = t_{\text{N-I},\alpha}\sqrt{\text{MQ}_e\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

Uma vez que o número de observações por grupo é o mesmo, a equação anterior poderá ser simplificada da seguinte forma:

$$|\overline{X}_1 - \overline{X}_2| > \text{LSD} = t_{N-I,\alpha}\sqrt{\text{MQ}_e\left(\frac{2}{N}\right)}$$

Este procedimento será repetido para todos os pares a comparar.

4.2.2 *Tukey's Honestly Significant Difference* (TSD)

Proposto por *Tukey* em 1953, este teste vem colmatar algumas lacunas verificadas nos restantes testes de comparação múltipla, nomeadamente na medida em que dele resultam intervalos de confiança menores.

Assim, é um dos testes de comparação de médias mais utilizado por ser bastante rigoroso mas também fácil de aplicar.

No caso das amostras serem de igual dimensão, o teste de Tukey (PortalAction, s.d.) considera que se rejeita a hipótese das médias serem iguais se:

$$|\overline{y}_i - \overline{y}_j| > \text{TSD} = q_\alpha(I, N-I)\sqrt{\frac{\text{MQe}}{n}},$$

onde q é um valor tabelado.

Tem-se ainda que um intervalo de confiança de 100(1-α)% para a diferença entre cada duas médias, é dado por:

$$\bar{y}_i - \bar{y}_j - q_\alpha(I, N-I)\sqrt{\frac{\text{MQe}}{n}} \leq \mu_i - \mu_j \leq \bar{y}_i - \bar{y}_j + q_\alpha(I, N-I)\sqrt{\frac{\text{MQe}}{n}}$$

4.3 Resultados

Neste ponto serão apresentados os resultados da análise de variância assim como dos testes *pairwise* para algumas das operações mencionadas no capítulo anterior.

Todos os testes serão efetuados para um nível de significância de 0,05.

4.3.1 *Downloads*

O número de *downloads* da aplicação, por trimestre, foram:

T1		T2		T3		T4	
Jan	6.132	Abr	5.634	Jul	3.704	Out	3.930
Fev	5.230	Mai	4.840	Ago	3.650	Nov	4.596
Mar	6.410	Jun	4.091	Set	3.394	Dez	3.501

Tabela 4.3: Número de *downloads* da aplicação por trimestre

Com base nos procedimentos descritos anteriormente na tabela 4.1, segue-se a construção da tabela ANOVA para os 4 grupos apresentados na tabela 4.3:

FONTE DE VARIAÇÃO	*SOMA DE QUADRADOS*	*GRAUS DE LIBERDADE*	*MÉDIA DE QUADRADOS*	*ESTATÍSTICA F*	*p-value*
Entre grupos	228.316.385,67	3	76.105.461,89	16,83	0,000812
Erro	36.171.588	8	4.521.448,50		
Total	264.487.973,67	11			

Tabela 4.4: ANOVA - *downloads*

Existem duas formas de interpretar a tabela anterior de modo a retirar conclusões sobre a mesma.

Vemos então que o *p-value* = 0,000812. Como este valor é inferior ao nível de significância considerado, rejeitamos a hipótese nula.

Para além disso, uma vez que a estatística F tem distribuição F com 3 e 8 graus de liberdade e, através da consulta da tabela F de Fisher-Snedecor a 5%, F(3,8) = 4,07, então a conclusão é a mesma e rejeitamos a hipótese nula.

Assim, conclui-se que existem evidências para afirmar que pelo menos um dos trimestres é diferente dos restantes.

Seguem-se os testes *pairwise* com o objetivo de averiguar quais são exatamente os trimestres que diferem dos outros.

Para tal, começa-se então por calcular o valor de LSD e TSD:

$$\text{LSD} = t_{N-I,\alpha}\sqrt{\text{MQe}\left(\frac{2}{N}\right)} = t_{8;0,05}\sqrt{\text{MQe}\left(\frac{2}{12}\right)} = 2,306\sqrt{4.521.448,5 \times \left(\frac{2}{12}\right)} = 4.003,62 \quad (4.1)$$

e

$$\text{TSD} = q_{\alpha}(I, N-I)\sqrt{\frac{\text{MQe}}{n}} = q_{0,05}(4,8)\sqrt{\frac{\text{MQe}}{3}} = 4,53\sqrt{\frac{4.521.448,5}{3}} = 5.561,30 \quad (4.2)$$

Em seguida, resta efetuar a diferença entre as médias de modo a concluir quais destas se encontram na região de rejeição de cada um dos testes.

	$\bar{X}_1$ = 21.767	$\bar{X}_2$ = 16.844	$\bar{X}_3$ = 11.416	$\bar{X}_4$ = 11.200
$\bar{X}_1$ = 21.767	0	4.923	10.352	10.567
$\bar{X}_2$ = 16.844		0	5.428	5.644
$\bar{X}_3$ = 11.416			0	215
$\bar{X}_4$ = 11.200				0

Tabela 4.5: Diferenças entre as médias das comparações *pairwise* - *downloads*

Analisando a tabela 4.5 e os valores obtidos em (4.1) e (4.2), pode observar-se que as conclusões diferem entre os dois testes.

No caso da aplicação do teste *Fisher's Least Significant Difference*, todas as diferenças entre médias à exceção de uma, são superiores ao valor de LSD, sendo que estas se encontram identificadas com o símbolo "*":

$$|\overline{X}_1 - \overline{X}_2| = 4.923 > \text{LSD} = 4.003,62^{*}$$

$$|\overline{X}_1 - \overline{X}_3| = 10.352 > \text{LSD} = 4.003,62\ ^*$$

$$|\overline{X}_1 - \overline{X}_4| = 10.567 > \text{LSD} = 4.003,62\ ^*$$

$$|\overline{X}_2 - \overline{X}_3| = 5.428 > \text{LSD} = 4.003,62\ ^*$$

$$|\overline{X}_2 - \overline{X}_4| = 5.644 > \text{LSD} = 4.003,62\ ^*$$

$$|\overline{X}_3 - \overline{X}_4| = 215 < \text{LSD} = 4.003,62$$

Isto significa que, à exceção das diferenças entre as médias dos trimestres 3 e 4, existem evidências para afirmar que todas as restantes sejam significativas uma vez que são superiores ao valor de LSD.

Conclui-se então que se consideram significativamente diferentes todos os pares de trimestres à exceção do terceiro e do quarto.

Por outro lado, para o teste *Tukey's Honestly Significant Difference* os resultados são um pouco distintos e, mais uma vez, assinalam-se com um "*"todas as diferenças consideradas significativas:

$$|\overline{X}_1 - \overline{X}_2| = 4.923 < \text{TSD} = 5.561,30$$

$$|\overline{X}_1 - \overline{X}_3| = 10.352 > \text{TSD} = 5.561,30\ ^*$$

$$|\overline{X}_1 - \overline{X}_4| = 10.567 > \text{TSD} = 5.561,30\ ^*$$

$$|\overline{X}_2 - \overline{X}_3| = 5.428 < \text{TSD} = 5.561,30$$

$$|\overline{X}_2 - \overline{X}_4| = 5.644 > \text{TSD} = 5.561,30\ ^*$$

$$|\overline{X}_3 - \overline{X}_4| = 215 < \text{TSD} = 5.561,30$$

Desta forma, considera-se que o trimestre 1 é significativamente diferente do trimestre 3 assim como do 4 e, além disso, existem também diferenças significativas entre os trimestres 2 e 4.

4.3.2 *Logins*

O número de *logins* na plataforma digital, por trimestre, foram:

T1		T2		T3		T4	
Jan	302 191	**Abr**	303 675	**Jul**	290 010	**Out**	297 939
Fev	295 534	**Mai**	308 906	**Ago**	290 287	**Nov**	317 731
Mar	315 072	**Jun**	300 245	**Set**	292 030	**Dez**	306 923

Tabela 4.6: Número de *logins* na plataforma digital por trimestre

Aplicando as fórmulas implícitas na tabela 4.1, vem que:

FONTE DE VARIAÇÃO	*SOMA DE QUADRADOS*	*GRAUS DE LIBERDADE*	*MÉDIA DE QUADRADOS*	*ESTATÍSTICA F*	*p-value*
Entre grupos	499.668.116,92	3	166.556.038,97	3,07	0,090976
Erro	434.184.666	8	54.273.083,25		
Total	933.852.782,92	11			

Tabela 4.7: ANOVA para *logins*

Ao contrário da análise dos *downloads*, vemos que *p-value* = 0,090976. Como este valor é superior ao nível de significância considerado, aceitamos a hipótese nula, ou seja, não existem diferenças entre os trimestres.

Também se pode verificar pela tabela F de Fisher-Snedecor a 5% que F(3,8) = 4,07 e, sendo este valor superior à estatística F, a conclusão será a mesma.

Assim, não será necessário recorrer aos testes *pairwise*, uma vez que se considera não haver diferenças significativas entre qualquer par de trimestres.

4.3.3 Consumos

O consumo médio mensal, em kW, do período em estudo foi:

T1		T2		T3		T4	
Jan	517	**Abr**	415	**Jul**	452	**Out**	480
Fev	498	**Mai**	401	**Ago**	487	**Nov**	502
Mar	475	**Jun**	425	**Set**	441	**Dez**	513

Tabela 4.8: Consumos médios por trimestre

Aplicando novamente os procedimentos descritos na tabela 4.1, vem que:

FONTE DE VARIAÇÃO	*SOMA DE QUADRADOS*	*GRAUS DE LIBERDADE*	*MÉDIA DE QUADRADOS*	*ESTATÍSTICA F*	*p-value*
Entre grupos	14.265,67	3	4.755,22	13,15	0,001852
Erro	2.894	8	361,75		
Total	17.159,67	11			

Tabela 4.9: ANOVA para consumos

Através da tabela 4.9 vemos que *p-value* = 0,001852. Como este valor é inferior ao nível de significância considerado, rejeitamos a hipótese nula.

Além disso, consultando a Tabela F de Fisher-Snedecor a 5%, o valor de F(3,8) = 4,07 é inferior à estatística F, assim rejeita-se a hipótese nula.

Desta forma, conclui-se que pelo menos um trimestre é diferente dos restantes e prossegue-se com a análise dos testes *pairwise* de modo a averiguar quais são exatamente os trimestres que diferem dos outros.

Assim, deverão ser calculados os valores de LSD e TSD:

$$\text{LSD} = t_{N-I,\alpha}\sqrt{\text{MQe}\left(\frac{2}{N}\right)} = t_{8;0,05}\sqrt{\text{MQe}\left(\frac{2}{12}\right)} = 2,306\sqrt{361,75\times\left(\frac{2}{12}\right)} = 35,81 \tag{4.3}$$

e

$$\text{TSD} = q_{\alpha}(I,N-I)\sqrt{\frac{\text{MQe}}{n}} = q_{0,05}(4,8)\sqrt{\frac{\text{MQe}}{3}} = 4,53\sqrt{\frac{361,75}{3}} = 49,74 \tag{4.4}$$

Para passar então à comparação entre os 4 trimestres, é necessário calcular as diferenças entre as médias de cada um dos pares:

Mais uma vez é possível observar através da figura 4.10 e dos valores obtidos em 4.3 e 4.4 que as conclusões obtidas em cada teste são distintas.

Começando pelo teste *Least Significant Difference*, verifica-se que existe apenas um par para o qual a diferença entre as suas médias é inferior ao valor de LSD. As diferenças significativas encontram-se representadas pelo símbolo ”*”:

	$\bar{X}_1$ = 497	$\bar{X}_2$ = 414	$\bar{X}_3$ = 460	$\bar{X}_4$ = 498
$\bar{X}_1$ = 497	0	83	37	2
$\bar{X}_2$ = 414		0	46	85
$\bar{X}_3$ = 460			0	38
$\bar{X}_4$ = 498				0

Tabela 4.10: Diferenças entre as médias das comparações *pairwise* - consumo

$$|\overline{X}_1 - \overline{X}_2| = 83 > \text{LSD} = 35,81^{\,*}$$

$$|\overline{X}_1 - \overline{X}_3| = 37 > \text{LSD} = 35,81^{\,*}$$

$$|\overline{X}_1 - \overline{X}_4| = 2 < \text{LSD} = 35,81$$

$$|\overline{X}_2 - \overline{X}_3| = 46 > \text{LSD} = 35,81^{\,*}$$

$$|\overline{X}_2 - \overline{X}_4| = 85 > \text{LSD} = 35,81^{\,*}$$

$$|\overline{X}_3 - \overline{X}_4| = 38 > \text{LSD} = 35,81^{\,*}$$

Isto leva à conclusão de que o único par de trimestres entre os quais não existem evidências para afirmar que hajam diferenças entre as suas médias são os trimestres 1 e 4.

No entanto, utilizando o teste *Tukey's Significant Difference*, as conclusões diferem comparativamente ao teste anterior. Os pares para os quais a diferença é significativa estão assinalados com o símbolo "*":

$$|\overline{X}_1 - \overline{X}_2| = 83 > \text{TSD} = 49,74^{\,*}$$

$$|\overline{X}_1 - \overline{X}_3| = 37 < \text{TSD} = 49,74$$

$$|\overline{X}_1 - \overline{X}_4| = 2 < \text{TSD} = 49,74$$

$$|\overline{X}_2 - \overline{X}_3| = 46 < \text{TSD} = 49,74$$

$$|\overline{X}_2 - \overline{X}_4| = 85 > \text{TSD} = 49,74^{\,*}$$

$$|\overline{X}_3 - \overline{X}_4| = 38 < \text{TSD} = 49,74$$

Neste caso, vemos que o trimestre 2 é significativamente diferente tanto do trimestre 1 como do trimestre 4.

4.3.4 Reclamações

O número de reclamações feitas tanto na plataforma digital como nos canais alternativos, por trimestre, foram:

T1		T2		T3		T4	
Jan	6 132	**Abr**	5 634	**Jul**	3 704	**Out**	3 930
Fev	5 230	**Mai**	4 840	**Ago**	3 650	**Nov**	4 596
Mar	6 410	**Jun**	4 091	**Set**	3 394	**Dez**	3 501

Tabela 4.11: Número de reclamações efetuadas por trimestre

Será então construída a tabela ANOVA, tendo como base os procedimentos descritos na tabela 4.1:

FONTE DE VARIAÇÃO	*SOMA DE QUADRADOS*	*GRAUS DE LIBERDADE*	*MÉDIA DE QUADRADOS*	*ESTATÍSTICA F*	*p-value*
Entre grupos	9.606.102,00	3	3.202.034,00	9,79	0,004699
Erro	2.615.583	8	326.947,83		
Total	12.221.684,67	11			

Tabela 4.12: ANOVA para reclamações

Conforme se pode verificar na figura 4.12, *p-value* = 0,004699. Uma vez que este valor é inferior ao nível de significância considerado, rejeitamos a hipótese nula.

Esta análise poderia ainda ser feita através da consulta da Tabela F de Fisher-Snedecor a 5%. Como a estatística F tem distribuição F com 3 e 8 graus de liberdade e F(3,8) = 4,07, então rejeitamos a hipótese nula.

Desta forma, conclui-se que pelo menos um trimestre é diferente dos restantes.

Assim, seguem-se os testes *pairwise* com o objetivo de averiguar quais são exatamente os trimestres que diferem dos outros.

$$\text{LSD} = t_{N-I,\alpha}\sqrt{\text{MQe}\left(\frac{2}{N}\right)} = t_{8;0,05}\sqrt{\text{MQe}\left(\frac{2}{12}\right)} = 2,306\sqrt{326.947,83 \times \left(\frac{2}{12}\right)} = 1.076,60 \quad (4.5)$$

e

$$\text{TSD} = q_{\alpha}(I,N-I)\sqrt{\frac{\text{MQe}}{n}} = q_{0,05}(4,8)\sqrt{\frac{\text{MQe}}{3}} = 4,53\sqrt{\frac{326.947,83}{3}} = 1.495,47 \quad (4.6)$$

Resta então calcular as diferenças entre as médias de todos os pares a comparar:

	$\bar{X}_1$ = 5.924	$\bar{X}_2$ = 4.855	$\bar{X}_3$ = 3.583	$\bar{X}_4$ = 4.009
$\bar{X}_1$ = 5.924	0	1.069	2.341	1.915
$\bar{X}_2$ = 4.855		0	1.272	846
$\bar{X}_3$ = 3.583			0	426
$\bar{X}_4$ = 4.009				0

Tabela 4.13: Diferenças entre as médias das comparações *pairwise* - reclamações

Por observação da figura 4.13 e dos valores obtidos em 4.5 e 4.6, é possível verificar que as conclusões obtidas diferem entre os dois testes.

O teste *Least Significant Difference* foi o primeiro a ser utilizado e verifica-se que existe apenas um par para o qual a diferença entre as suas médias é inferior ao valor de LSD. Todos os restantes, ou seja, para os quais existem evidências para se afirmar que as suas médias são diferentes, representam-se pelo símbolo "*":

$$|\overline{X}_1 - \overline{X}_2| = 1.069 < \text{LSD} = 1.076,70$$

$$|\overline{X}_1 - \overline{X}_3| = 2.341 > \text{LSD} = 1.076,70^{*}$$

$$|\overline{X}_1 - \overline{X}_4| = 1.915 > \text{LSD} = 1.076,70^{*}$$

$$|\overline{X}_2 - \overline{X}_3| = 1.272 > \text{LSD} = 1.076,70^{*}$$

$$|\overline{X}_2 - \overline{X}_4| = 846 < \text{LSD} = 1.076,70$$

$$|\overline{X}_3 - \overline{X}_4| = 426 < \text{LSD} = 1.076,70$$

Isto leva à conclusão de que há evidências para afirmar que existem diferenças entre as médias dos trimestres 2 e 3 e, além desses, também o trimestre 1 é significativamente diferente do trimestre 3 e do 4.

No entanto, utilizando o teste *Tukey's Significant Difference*, as conclusões diferem comparativamente ao teste anterior. Os pares significativos assinalam-se com o símbolo "*":

$$|\overline{X}_1 - \overline{X}_2| = 1.069 < \text{TSD} = 1.495,47$$

$$|\overline{X}_1 - \overline{X}_3| = 2.341 > \text{TSD} = 1.495,47^{\,*}$$

$$|\overline{X}_1 - \overline{X}_4| = 1.915 > \text{TSD} = 1.495,47^{\,*}$$

$$|\overline{X}_2 - \overline{X}_3| = 1.272 < \text{TSD} = 1.495,47$$

$$|\overline{X}_2 - \overline{X}_4| = 846 < \text{TSD} = 1.495,47$$

$$|\overline{X}_3 - \overline{X}_4| = 426 < \text{TSD} = 1.495,47$$

Neste caso, conclui-se que existem evidências para afirmar que o trimestre 1 é significativamente diferente dos trimestres 3 e 4.

Capítulo 5

Conclusão

No que diz respeito à análise de *clusters*, esta permitiu identificar o tipo de perfil de um cliente digital da empresa. Esta análise foi facilitada pelo agrupamento das características, com base no seu consumo, faixa etária e região.

Através do método de *k-means*, foi possível concluir que o cliente digital comum é essencialmente um indivíduo pertencente a uma faixa etária mais baixa, nomeadamente até aos 30 anos, e com um consumo mensal mais alto.

Assim, fará mais sentido apostar em campanhas massivas direcionadas a um público numa faixa etária superior, uma vez que os outros ou já se encontram vinculados ou serão facilmente convertidos ao digital. No entanto, são os consumidores mais jovens que têm um comportamento mais oscilante, digamos assim. De um dia para o outro, facilmente mudam de opinião e isso pode ser um risco para a empresa, devendo esta precaver-se dessa situação também através de campanhas específicas para este segmento de clientes.

Relativamente ao comportamento do cliente na plataforma digital face aos outros canais alternativos, há ainda um longo caminho a percorrer pois a diferença entre ambos é ainda muito significativa.

Pela análise de sazonalidade, foi possível identificar um padrão em quase todos os testes efetuados. Apesar de terem sido obtidos resultados distintos em cada um, través dos testes *Least Significant Difference* e *Tukey's Significant Difference* verificou-se que há uma tendência para que o comportamento do cliente se altere durante os meses de verão.

Numa futura abordagem ao tema, seria interessante calcular a probabilidade de um cliente se tornar digital com base nas suas características e comportamento histórico assim como nas de outros clientes que se tornaram digitais ao longo do tempo.

Bibliografia

Alpuim, T. (2017). *Apontamentos da Cadeira de Modelos Lineares*.

DataNovia (s.d.). *K-Means Clustering in R: Algorithm and Practical Examples*. URL: `https://www.datanovia.com/en/lessons/k-means-clustering-in-r-algorith-and-practical-examples/`.

MacQueen, James et al. (1967). "Some methods for classification and analysis of multivariate observations". Em: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*.

PortalAction (s.d.). *Teste de Tukey*. URL: `http://www.portalaction.com.br/anova/31-teste-de-tukey`.

Tan, Pang-Ning, Michael Steinbach, Vipin Kumar et al. (2006). "Cluster analysis: basic concepts and algorithms". Em: *Introduction to data mining*.

Williams, Lynne J e Herve Abdi (2010). "Fisher's least significant difference (LSD) test". Em: *Encyclopedia of research design*.

Anexos

Lista de variáveis consideradas:

Região

Alentejo: se o indivíduo pertence à região do Alentejo;

Algarve: se o indivíduo pertence à região do Algarve;

Beira: se o indivíduo pertence à região da Beira;

Estremadura: se o indivíduo pertence à região da Estremadura;

Minho: se o indivíduo pertence à região do Minho;

Trás os Montes: se o indivíduo pertence à região de Trás os Montes;

Idade

Até 30: se o indivíduo tem até 30 anos;

Entre 30 e 45: se o indivíduo tem entre 30 e 45 anos;

Entre 45 e 60: se o indivíduo tem entre 45 e 60 anos;

Superor a 60: se o indivíduo tem mais de 60 anos;

Consumo

Muito baixo: 1 se o indivíduo tem um consumo médio de até 150 kw por mês;

Baixo: 2 se o indivíduo tem um consumo médio de 150 a 250 kw por mês;

Médio: 3 se o indivíduo tem um consumo médio de 250 a 450 kw por mês;

Alto: 4 se o indivíduo tem um consumo médio de 450 a 600 kw por mês;

Muito alto: 5 se o indivíduo tem um consumo médio superior a 600 kw por mês;