UNIVERSIDADE DE SÃO PAULO

ESCOLA DE ENGENHARIA DE LORENA

MARINA SERVO MORALES

Perfis de clientes do mercado financeiro de adquirência por meio de análise de banco de dados

Lorena

2021

MARINA SERVO MORALES

Perfis de clientes do mercado financeiro de adquirência por meio de análise de banco de dados

Trabalho de Conclusão de Curso apresentado à Escola de Engenharia de Lorena da Universidade de São Paulo como requisito para a obtenção do título de engenheira física

Orientador: Prof. Dr. Carlos Yujiro Shigue

Lorena

2021

Ficha catalográfica elaborada pelo Sistema Automatizado
da Escola de Engenharia de Lorena,
com os dados fornecidos pelo(a) autor(a)

Morales, Marina Servo
Perfis de clientes do mercado financeiro de adquirência por meio de análise de banco de dados / Marina Servo Morales; orientador Carlos Yujiro Shigue. – Lorena, 2021.
49 p.

Monografia apresentada como requisito parcial para a conclusão de Graduação do Curso de Engenharia Física – Escola de Engenharia de Lorena da Universidade de São Paulo. 2021

1. Persona. 2. K-means. 3. Sql. 4. Rstudio. I. Título. II. Shigue, Carlos Yujiro, orient.

Dedico este trabalho aos meus pais, sem eles, nada disso seria possível.

AGRADECIMENTOS

Agradeço a Escola de Engenharia de Lorena da Universidade de São Paulo por todo aprendizado e oportunidades que a graduação me proporcionou. Agradeço a todos os professores que passaram pelo meu caminho e a todos os ensinamentos que me deram, em especial ao professor Carlos Yujiro Shigue, pela sua orientação neste trabalho.

Agradeço a todas as pessoas que passaram pelo meu caminho na minha vida profissional, me ensinando, motivando e me inspirando, pois estes foram peças fundamentais na motivação deste estudo.

Agradeço ao meu namorado, Johan Lemes dos Santos, que me apoiou e me motivou em todos os anos de graduação, inclusive nesta etapa final do trabalho de graduação.

Agradeço a Deus e a minha família, principalmente aos meus pais, Alessandra Servo Morales e Carlos Martin Morales, que me apoiaram e me deram forças em todos os momentos, desde o vestibular, até esta etapa final. Fizeram o possível e o impossível para que eu conseguisse cursar a graduação e realizar este sonho.

“A dúvida é o princípio da sabedoria”.

Aristóteles

RESUMO

MORALES, Marina Servo. **Perfis de clientes do mercado financeiro de adquirência por meio de análise de banco de dados.** 2021. Número de folhas 49f. Monografia (Graduação) – Escola de Engenharia de Lorena, Universidade de São Paulo, Lorena, 2021.

Grandes empresas têm uma grande e diversa quantidade de dados nos tempos atuais. Entretanto, ainda assim, estes dados não têm todo seu potencial aproveitado, muitas vezes os clientes não têm suas dores compreendidas e as empresas não conseguem oferecer as soluções que seus clientes precisam. O objetivo central do trabalho é fazer a extração e análise dos dados dos clientes de uma empresa brasileira financeira que presta serviços de adquirência para otimizar o direcionamento das campanhas de marketing e oferta de soluções para seus clientes. Propõe-se, assim, fazer a extração da base por meio da linguagem SQL na interface gráfica HUE e a análise de agrupamentos por meio do *software* estatístico RStudio, utilizando os métodos de agrupamento hierárquico Ward e não-hierárquico K-*means* e, por fim, fazer a criação das *personas* destes clientes. Os resultados obtidos dos cinco agrupamentos e personas criados foram satisfatórios e a empresa poderá utilizá-los como direcionadores em suas ações futuras.

Palavras-chave: Persona, K-means, SQL, RStudio.

ABSTRACT

MORALES, Marina Servo. **Acquiring financial market customer profiles through database analysis.** 2021. Number of sheets 49. Monograph (Graduation) – Engineering School of Lorena, University of São Paulo, Lorena, 2021.

Large companies have a large and diverse amount of data these days. However, still this data does not have its full potential, customers often do not have their pains understood and companies are unable to offer the solutions their customers need. The main objective of this work is to extract and analyze customer data from a Brazilian financial company that provides acquiring services to optimize the targeting of marketing campaigns and offer solutions to its customers. It is proposed, therefore, to extract the base through the SQL language in the HUE graphical interface and to make the analysis of clusters through the statistical software RStudio, using the hierarchical clustering methods Ward and non-hierarchical K-means, and finally, create the personas of these customers. The results obtained from the five groups and personas created were satisfactory and the company will be able to use them as guidelines in its future actions.

Keywords: Persona, K-means, SQL, RStudio.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE SIGLAS

HUE	*Hadoop User Experience*
POS	*Point of Sale*
SQL	*Structured Query Language*

LISTA DE SÍMBOLOS

Σ	Somatória
EU	Distância Euclidiana
i	Dimensão com n=1 espaço euclidiano
n	n dimensões do espaço euclidiano

SUMÁRIO

1 INTRODUÇÃO

No cenário atual, grandes empresas, com uma grande quantidade de clientes, estão bem munidas de todo tipo de informação. O banco de dados é uma ótima ferramenta para tais casos, pois torna acessível essa grande quantidade de dados para todos os colaboradores da empresa. Contudo, essa abundância de informações muitas vezes é sub aproveitada, pois são feitas sempre as mesmas bases e as mesmas análise, isso se deve tanto por falta de conhecimento do potencial desses dados, quanto por falta de treinamento por parte da empresa.

Outro fator importante é que, muitas vezes, as empresas não tem conhecimento de fato de qual é a real "dor do cliente", oferecem soluções que acreditam ser o que o cliente necessita e, por este motivo, alguns produtos são descontinuados pouco tempo depois que são lançados pela empresa. Não ter conhecimento das personas pode gerar o que é conhecido como o "usuário elástico", que é a definição do usuário segundo as percepções pessoais que cada colaborar tem e, com isso, gerando um usuário que se adapta ao conceito de cada um (BARROS, 2019).

Apesar dos métodos mais utilizados para criação das *personas* ainda serem, entrevistas e questionários, existem metodologias e *softwares* que podem facilitar e otimizar esse processo. Um exemplo de metodologia é a Análise de Agrupamentos, que une os dados semelhantes em grupos homogêneos. As três etapas do processo de Análise de Agrupamentos são, primeiramente, definição do método de cálculo de similaridade entre os grupos, qual método para formação dos grupos e, por fim, quantos grupos serão formado (SEIDEL *et al*., 2008).

O método mais utilizado para o cálculo da similaridade entre os grupos é o das distâncias euclidianas e este é o utilizado neste trabalho. Os métodos de formação dos agrupamentos dividem-se entre métodos hierárquicos e não-hierárquicos, neste trabalho serão utilizados o método hierárquico de Ward para a definição da melhor quantidade de agrupamentos e o método não-hierárquico de K-*means* para a formação dos agrupamentos.

O método de Ward tende a gerar agrupamentos de tamanhos similares por causa do cálculo de minimização da variação interna que realiza (SEIDEL *et al*., 2008). Já o método de K-means, segundo Aquino Neto e Araujo (2020), é muito utilizado para realizar agrupamentos em grandes bases de dados.

O estudo para criação das *personas* tem como objetivo entender o perfil dos clientes de uma empresa financeira brasileira de adquirência para, então, otimizar a oferta de campanhas de marketing e oferta de soluções para seus clientes.

2 OBJETIVO

O presente trabalho teve como objetivo fazer a extração de uma base de dados de uma empresa financeira brasileira prestadora de serviços de adquirência, com mais de 1,2 milhão de clientes ativos, através da linguagem de programação SQL na interface gráfica HUE, e a análise de agrupamentos destes dados, utilizando os métodos de Ward e K-means, possibilitando então, a criação de *personas* com estes agrupamentos para otimizar as campanhas de marketing e ofertas de produtos aos clientes.

3 FUNDAMENTAÇÃO TEÓRICA

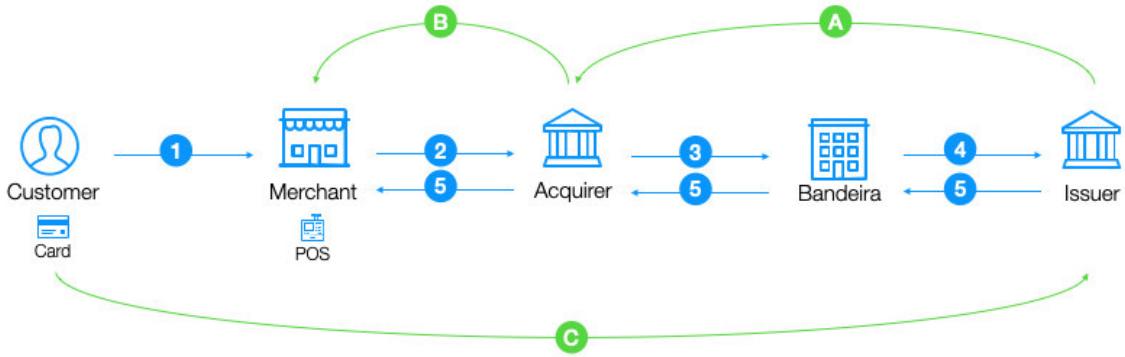
A seguir está a fundamentação teórica referente aos temas abordados no presente trabalho.

3.1 Mercado Financeiro de Adquirência

As empresas de adquirências, também conhecidas como credenciadoras, são as responsáveis pelo credenciamento dos estabelecimentos comerciais para realização das transações nas máquinas de cartão, chamadas de POS (Point of Sale), tradução de *Point of Sale*. Também são as responsáveis pela transmissão dos dados das transações do POS para a bandeira, como Visa e Mastercard, e, por fim, receber a resposta da transação recebida do banco emissor e da bandeira (ALMEIDA, 2013).

Na figura 1, está representado o fluxograma de uma transação no POS.

Figura 1. Fluxo de transação do estabelecimento comercial ao banco emissor



Fonte: (ROSA, 2020)

3.2 *Persona*

Segundo Barros (2019), a técnica de *personas* foi criada em 1995 pelo especialista de marketing Angus Jenkinson e pelo especialista de desenvolvimento de software Alan Cooper e tem como definição “arquétipos hipotéticos dos usuários atuais”.

Com as *personas*, entende-se o comportamento e objetivos de um grupo de pessoas em situações específicas. Tradicionalmente, as ferramentas mais utilizadas no desenvolvimento das *personas* são os questionários, entrevistas, designs etnográficos, *focus group, card sorting*, testes de usabilidade e análises de tarefa (BARROS, 2019). Entretanto, existem ferramentas estatísticas que possibilitam a criação das *personas* através de algoritmos de análise de agrupamento, como por exemplo, o *software* R.

No fluxograma da figura 2, está apresentada as etapas do processo de criação de *personas*.

Figura 2. Fluxograma com as sete etapas de construção de *personas*

Fonte: Adaptado de Barros (2019)

3.3 Análise de Agrupamentos

Segundo Moori, Marcondes e Ávila (2006, p. 71):

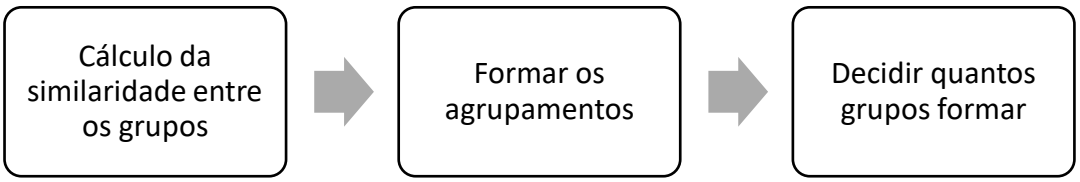
> A análise de agrupamentos (cluster analysis) é uma técnica estatística que permite ao pesquisador separar ou classificar objetos observados em um grupo ou em número específico de subgrupos ou conglomerados (clusters) mutuamente exclusivos, de modo que os subgrupos formados tenham características de grande similaridade interna e grande dissimilaridade externa.

Aglomeração hierárquica e não-hierárquica são dois tipos de análises de agrupamentos em que, na aglomeração hierárquica, como o próprio nome sugere, forma-se uma hierarquia,

ou uma árvore, com os objetos semelhantes entre si, sua análise gráfica é feita através dos dendogramas e a quantidade de agrupamentos é definida após sua análise. Um método de análise de agrupamento hierárquico é o Método de Ward. Na aglomeração não-hierárquica, o método utilizado é o K-*means* e nele formam-se os agrupamentos a partir de um centro de agrupamento. Estes agrupamentos estão à uma distância predeterminada desse agrupamento central e a quantidade de agrupamentos, ou centroides, é definida após sua análise (SEIDEL *et al*., 2008).

É possível ver as etapas da análise de agrupamentos na figura 3.

Figura 3. Diagrama das três etapas da Análise de Agrupamentos



Fonte: Adaptado de Seidel *et al*. (2008)

O primeiro passo da análise de agrupamentos é fazer o cálculo da similaridade entre os grupos, ou seja, o cálculo das distâncias de cada grupo com todos os outros grupos, que neste trabalho será feito pelas distâncias euclidianas entre as variáveis estudadas. O cálculo pelas distâncias euclidianas é o método mais utilizado quando as variáveis são quantitativas (MOORI, MARCONDES, ÁVILA, 2006).

A distância euclidiana entre dois pontos que possuem n dimensões é calculada pela equação 1 (AQUINO NETO; ARAUJO, 2020).

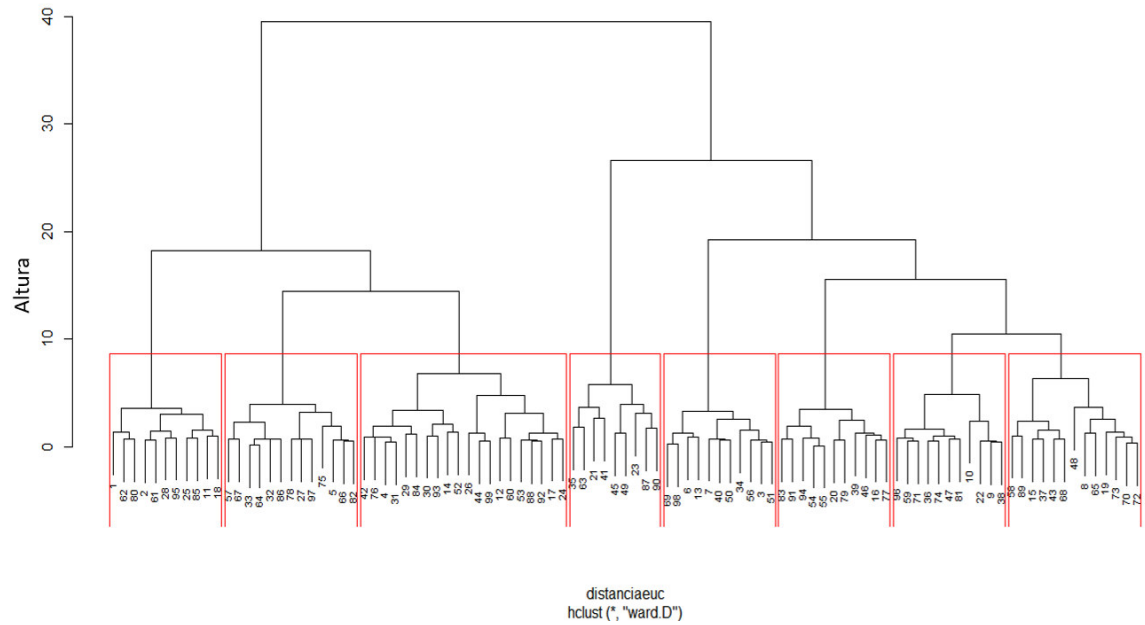
$$EU\ (x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (1)$$

3.3.1 Ward

Também conhecido como "Mínima Variância", o método hierárquico de Ward foi desenvolvido em 1963 pelo professor Ward da Universidade de Stanford. Este método tem como primeiro passo, como mencionado anteriormente, o cálculo da similaridade dos grupos pela distância euclidiana entre dois pontos (MINGOTI, 2005).

Feito isso, os grupos similares são agrupados e definidos de acordo com o dendograma formado, como o da figura 4 (SEIDEL *et al*., 2008).

Figura 4. Exemplificação de um dendograma com os agrupamentos realizados pelo Método de Ward



Fonte: Própria autora

Na exemplificação da figura 4, foram definidos 8 agrupamentos de acordo com o dendograma.

No software R, existem dois algoritmos para o agrupamento por Ward, o “ward.d” e o “ward.D2” e a diferença entre eles é que, no segundo algoritmo as dissimilaridades são elevadas ao quadrado antes da atualização do agrupamento (R Core Team, 2021).

3.3.2 Single-link

O Single-link é um método hierárquico onde os dois indivíduos mais próximo se unem e formam o primeiro agrupamento. Em seguida, de acordo com o critério da distância mínima, um terceiro indivíduo pode se unir ou primeiro agrupamento formado, ou um novo agrupamento pode ser formado entre dois indivíduos que têm uma distância menor entre si (VALLI, 2012).

3.3.3 Complete linkage

Este método hierárquico é idêntico ao single-link, com exceção de que o critério de agrupamento é a maior distância entre dois indivíduos, e não a menor (VALLI, 2012).

3.3.4 Average linkage

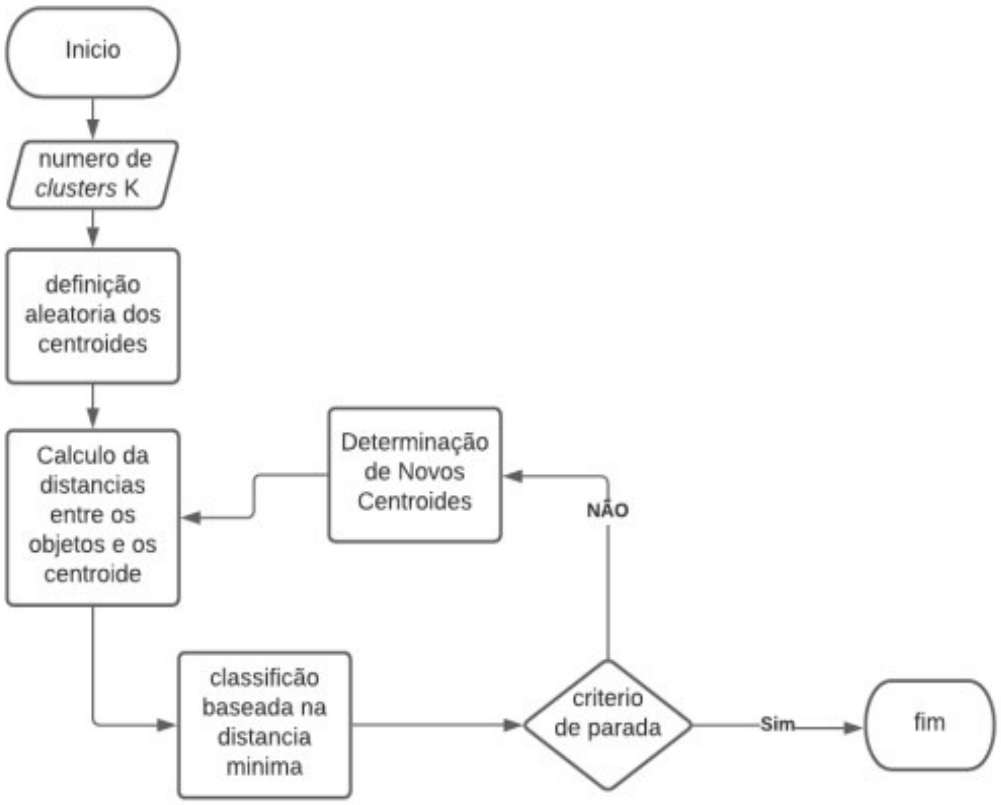
O método hierárquico average linkage segue o mesmo fluxo dos dois métodos anteriores, contudo o critério é a média entre dois indivíduos (VALLI, 2012).

3.3.5 K-*means*

Segundo Seidel *et al*. (2008), o método não-hierárquico de K-means é muito utilizado em estudos com muitos objetos e que têm uma pequena variação entre si.

Em cada iteração deste método, seu algoritmo faz o agrupamento de todos os dados de acordo com a distância entre cada um e cada centroide. Cada centroide é inicializado de forma aleatória no início da primeira iteração e, no fim de cada iteração são atualizados com os valores da média aritmética dos dados de seu agrupamento. O fim do processo ocorre quando não há mais mudanças significativas entre as iterações. O processo pode ser visualizado no fluxograma da figura 5.

Figura 5. Fluxograma do método de agrupamentos não-hierárquico K-*means*



Fonte: (AQUINO NETO; ARAUJO, 2020)

3.4 Banco de Dados

Banco de dados é um conjunto de tabelas em que cada uma armazena dados de um agrupamento específico. Essas tabelam se relacionam através de alguns campos e, com isso, facilitam a extração de bases complexas, com uma extensa quantidade de informações. Se necessário, dados sensíveis podem ter acesso mais restrito com tabelas que apenas alguns colaboradores de uma empresa tenham acesso, por exemplo (NIELD, 2016).

Existem dois tipos de bancos de dados, os relacionais e os não relacionais. O banco de dados relacional é aquele que está estruturado por tabelas e cada linha é um registro, que pode se relacionar com as diferentes colunas, ou seja, a tabela é a relação. Já o banco de dados não relacional, como o próprio nome sugere, não tem dados que se relacionam entre si na mesma tabela, um exemplo são as imagens presentes nas redes sociais (NISHI; SOUZA; SANTANA, 2017).

3.5 SQL

SQL é uma linguagem de programação onde seu nome vem de SEQUEL por ser uma sequência de outros protótipos de linguagens. Seu primeiro padrão de linguagem foi desenvolvido em 1980 e o último em 2006 segundo Nield (2016).

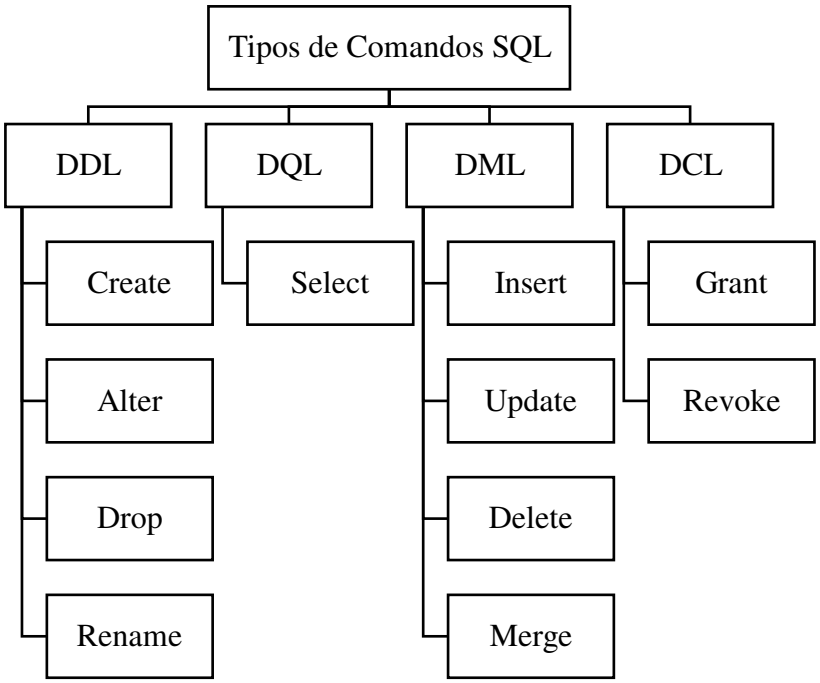
Uma query é o código de programação para extração dos dados dos bancos de dados e sua estrutura básica é composta do SELECT, do FROM e do WHERE, onde primeiramente é definido quais campos deseja-se no *output*, depois é definido qual ou quais tabelas são usadas para extração dos campos definidos e, por fim, os filtros desejados, como período, lugar e etc, são determinados, respectivamente (NIELD, 2016).

Outras funções tão importantes quanto são o CREATE TABLE, DROP TABLE e os JOIN's. Na seção do FROM existem diferentes tipos de JOIN, existe o LEFT, o RIGHT, o INNER e o FULL JOIN (NIELD, 2016).

A linguagem SQL está dividida em quatro tipos de comando, o DDL, ou "Data Definition Language", o DQL, ou "Data Query Language", o DML, ou "Data Manipulation Language" e o DCL, ou "Data Control Language" (SILVA, 2019).

Estes comandos estão descritos na figura 6.

Figura 6. Divisão dos tipos de comandos no SQL: DDL, DQL, DML e DCL



Fonte: (SILVA, 2019)

3.5.1 SELECT e CREATE TABLE

O CREATE TABLE é utilizado para criação de tabelas no banco de dados, mas não é necessário criar uma tabela em cada extração de base, somente quando se deseja que a tabela fique definitivamente no banco de dados.

Já o SELECT é a estrutura inicial obrigatória de uma query, onde é definido quais campos das tabelas serão utilizados no *output*, ou seja, quais serão selecionados. Este comando é seguido do CREATE TABLE, quando utilizado (NIELD, 2016).

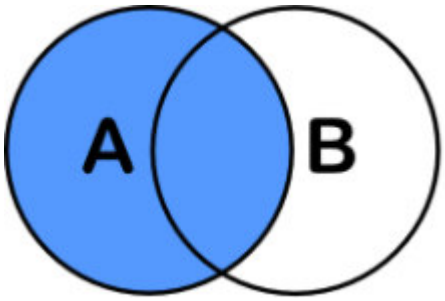
3.5.2 DROP TABLE

O DROP TABLE é necessário quando se deseja excluir uma tabela do banco dados (NIELD, 2016).

3.5.3 LEFT JOIN

O LEFT JOIN é o comanda complementar mais utilizado do FROM porque possibilita extrair os dados complementares da tabela B a partir dos dados primários definidos da tabela A, ou seja, retorna todas as linhas da tabela A e os campos da tabela B que fazem interseção com os campos chave definidos na query, como mostra a figura 7 (NIELD, 2016).

Figura 7. Diagrama da representação das tabelas em um LEFT JOIN

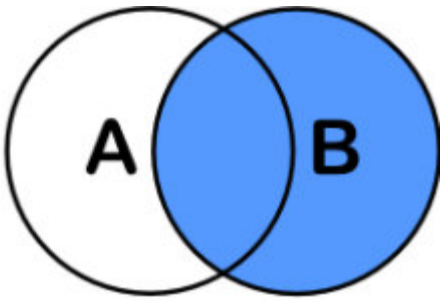


Fonte: (FABIOBMED, 2013)

3.5.4 RIGHT JOIN

Já o RIGHT JOIN, ao contrário do LEFT JOIN, retorna todas as linhas que têm interseção dos campos entre as tabelas A e B e as linhas restantes da tabela B, que não fizeram interseção com a tabela A. Pode-se ter uma visualização do conceito do RIGHT JOIN na figura 8 (NIELD, 2016).

Figura 8. Diagrama da representação das tabelas em um RIGHT JOIN

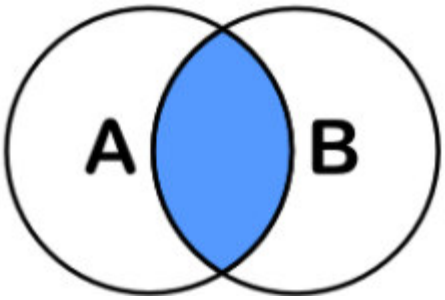


Fonte: (FABIOBMED, 2013)

3.5.5 INNER JOIN

O INNER JOIN é representado na figura 9 (NIELD, 2016).

Figura 9. Diagrama da representação das tabelas em um INNER JOIN



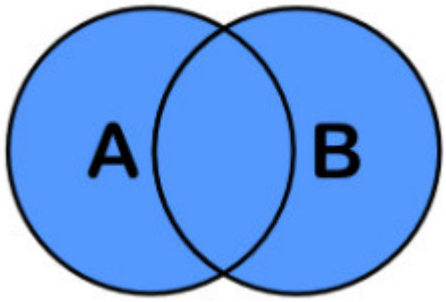
Fonte: (FABIOBMED, 2013)

Como se pode ver pela imagem, o INNER JOIN retorna apenas as linhas em que há interseção dos campos chave entre a tabela A e B.

3.5.6 FULL JOIN

Por fim, o FULL JOIN é utilizado para retornar todas as linhas de ambas tabelas, A e B, como mostra a figura 10 (NIELD, 2016).

Figura 10. Diagrama da representação das tabelas em um FULL JOIN



Fonte: (FABIOBMED, 2013)

3.5.7 GRANT e REVOKE

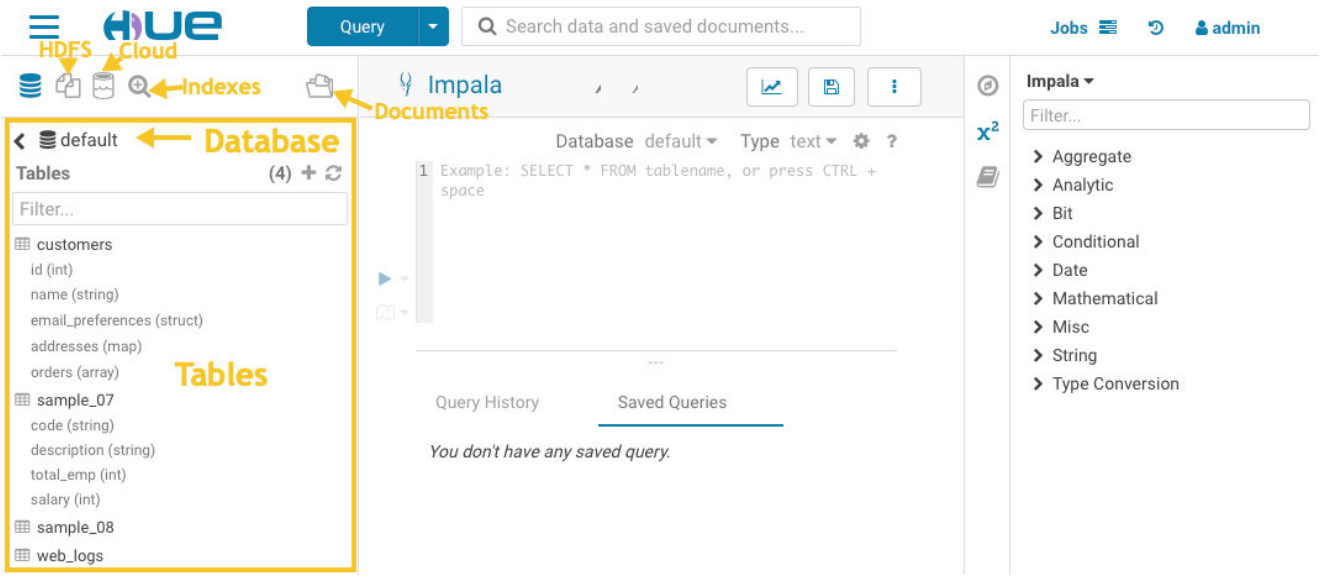
O comando GRANT é utilizado para conceder acesso a uma determinada tabela ao usuário especificado ou a utilizar certos comandos em uma tabela, como o DELETE, SELECT entre outros. Já o comando REVOKE é utilizado para revogar os acessos concedidos ao usuário a tabelas ou a realizar certos comandos nas tabelas que ele tenha acesso (SILVA, 2019).

3.6 Apache Hadoop e HUE

O Apache Hadoop é uma biblioteca de software, com um a milhares de servidores, que possibilita a usabilidade de grandes agrupamentos de dados. O Hadoop é da empresa Apache Software Foundation, fundada em 1999, uma empresa sem fins lucrativos que tem como objetivo apoiar os softwares de código aberto (APACHE, 2021).

O HUE, ou Hadoop User Experience, é o mais conhecido e popular sistema de gestão de *Big Data*, agrupando diversos projetos do Apache Hadoop. Na figura 11 é possível ver a interface de programação do HUE e seus detalhes (AWS, 2021).

Figura 11. Interface HUE



Fonte: (CLOUDERA, 2021)

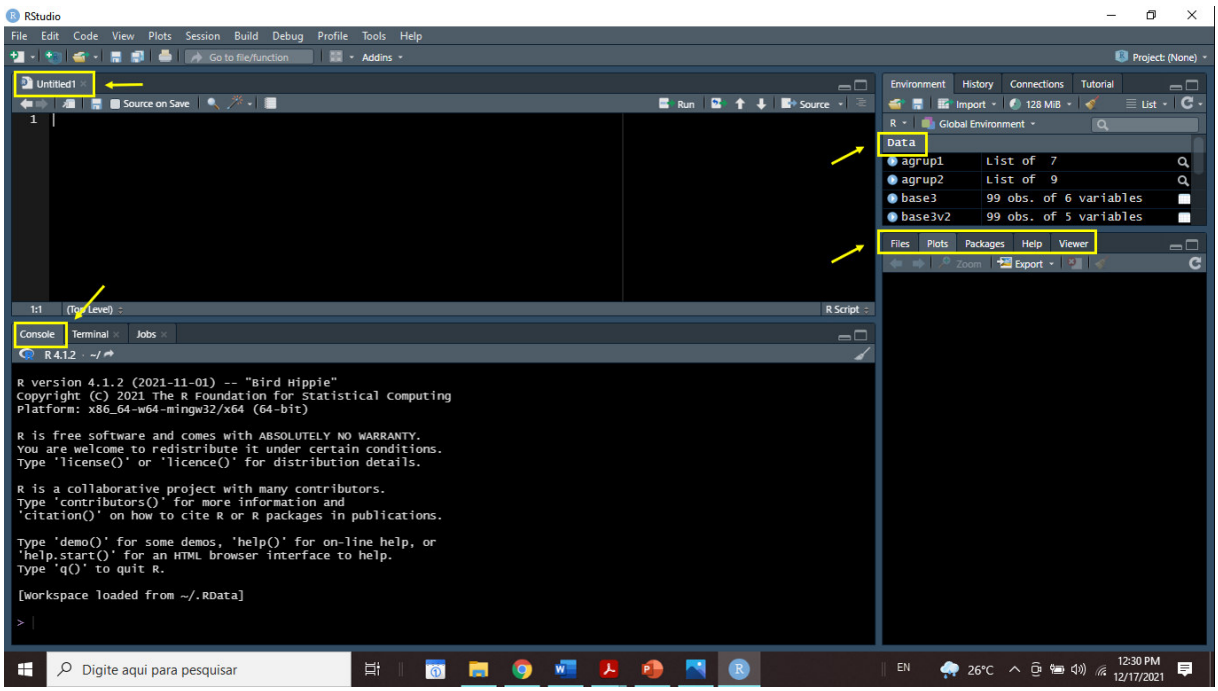
3.7 *Software* R

O R é um *software* computacional estatístico criado em 1996 por professores de estatística da Universidade de Auckland. De código aberto, grátis e GNU (General Public License), o R permite que seus usuários modifiquem seu código-fonte de acordo com a sua necessidade. É um software que funciona a partir de diversos "pacotes" e que qualquer pessoa pode desenvolver um "pacote" para a aplicação que desejar. Sua versão básica tem instalado cerca de 8 pacotes essenciais, o que o torna um *software* rápido e leve. Contudo, é de extrema facilidade a instalação de outros pacotes necessários, apenas com a função "install.packages("")".

Existe também o ambiente integrada do R chamado RStudio, igualmente gratuito, sendo este um ambiente mais amistoso, com janelas de plotagem para facilitar a visualização dos resultados (RITTER et al., 2019).

Na figura 12, é possível ver a interface de programa do RStudio, onde existe uma área, no canto superior esquerdo, para criação dos códigos de programação em R, uma área chamada "Console" onde é possível compilar as linhas de código, a área "Data" onde as variáveis e tabelas criadas ficam salvas e existe a área com os arquivos e *plots* criados, no canto inferior direito (RITTER et al., 2019).

Figura 12. Interface de programação do *software* RStudio



Fonte: Própria autora

4 MATERIAIS E MÉTODOS

Nesta seção são descritos as variáveis, os softwares e os métodos utilizados para as análises deste estudo. O processo de criação das *personas* foi adaptado para ser condizente com a realidade do estudo, não foram realizadas entrevistas, entretanto, as variáveis relevantes foram identificadas e extraídas a partir do banco de dados da empresa em questão.

4.1 Variáveis utilizadas para análise

As variáveis utilizadas para a análise do estudo foram definidas com base no objetivo de criação das *personas*, ou seja, são variáveis relevantes para a empresa no direcionamento de campanhas de marketing e oferecimento de soluções.

A seguir, está o quadro 1 com o "de para" das variáveis fictícias que foram criadas para preservação dos dados da empresa em estudo.

Quadro 1 - Variáveis utilizadas na query para extração da base utilizada para o estudo

Variável	**Descrição da variável**
id_estab_comercial	Número de identificação do estabelecimento comercial
data_cadastro	Data em que o estabelecimento comercial se cadastrou na empresa adquirente
status_estab_comercial	Status de atividade do estabelecimento comercial (ATIVO = transacionou a menos de 30 dias, INATIVO = não transaciona a mais de 30 dias)
faturamento	Valor bruto de venda autorizada
qtde_vendas	Quantidade de vendas sobre o faturamento
ano_mes	Código do valor 100*ano+mês para identificação do período da venda
nome_origem	Origem do dado: app / site
qtde_acessos	Quantidade de acessos no app ou site no período definido

Fonte: Própria autora

4.2 Query em SQL

A seguir está a *query* SQL Impala, criada na interface gráfica Hue, para extração de uma amostra da base de 1,87 milhões de clientes ativos, do dia 1 de dezembro de 2021, com período de faturamento de quatro meses, de agosto de 2021 a novembro de 2021.

A amostra utilizada é uma base de 25.000 clientes ativos da empresa em estudo. O tamanho da amostra foi definido pela limitação computacional do equipamento utilizado.

Abaixo segue a descrição da query utilizada.

```
WITH BASE_CADASTRO as (
SELECT
     id_estab_comercial
     ,months_between(now(),data_cadastro) as MOB_meses_202112

FROM cadastro.tabela_cadastro

WHERE status_estab_comercial = 'ATIVO'),

FAT_MENSAL as (
SELECT
     A.id_estab_comercial
     ,SUM(B.faturamento)/4 AS MD_FAT
     ,SUM(B.qtde_vendas)/4 AS MD_VENDAS

FROM BASE_CADASTRO as A
LEFT    JOIN    faturamento.tabela_faturamento    as    B    ON
(A.id_estab_comercial= B.id_estab_comercial)

WHERE B.ano_mes in (202108,202109,202110,202111)

GROUP BY 1),

BASE_SITE_APP as (
SELECT
     A.id_estab_comercial
     ,SUM(CASE  WHEN  C.nome_origem="app"  THEN  C.qtde_acessos
ELSE 0 END)/4 as MD_ACESSOS_APP
     ,SUM(CASE WHEN C.nome_origem="site" THEN C.qtde_acessos
ELSE 0 END)/4 as MD_ACESSOS_SITE

FROM BASE_CADASTRO as A
LEFT    JOIN    faturamento.tabela_acessos    as    C    ON
(A.id_estab_comercial= C.id_estab_comercial)

WHERE C. ano_mes in (202108,202109,202110,202111)

GROUP BY 1)
```

```
SELECT
A.*
,D. MD_FAT
,D.MD_VENDAS
,F. MD_ACESSOS_APP
,F. MD_ACESSOS_SITE

FROM BASE_CADASTRO as A
LEFT JOIN FAT_MENSAL as D ON (A.id_estab_comercial=
D.id_estab_comercial)
LEFT JOIN BASE_SITE_APP as F ON (A.id_estab_comercial=
F.id_estab_comercial)

ORDER BY A.id_estab_comercial
```

No quadro 2, está o “de para” das variáveis criadas na query.

Quadro 2 - Variáveis criadas na query para utilização na análise de agrupamentos

Variável	**Descrição da variável**
MOB_meses_202112	Tempo de empresa do estabelecimento comercial
MD_FAT	Média de faturamento de agosto a novembro de 2021 do estabelecimento comercial
MD_VENDAS	Média de quantidade de vendas de agosto a novembro de 2021 do estabelecimento comercial
MD_ACESSOS_APP	Média de acessos de agosto a novembro de 2021 no aplicativo de gestão da empresa
MD_ACESSOS_SITE	Média de acessos de agosto a novembro de 2021 no site de gestão da empresa

Fonte: Própria autora

A base foi extraída e as análises foram feitas no notebook da empresa em estudo, contudo, devido a Lei Geral de Proteção de Dados, nenhum arquivo pode ser enviado para e-mails externos e a base utilizada neste trabalho foi uma simulação com as médias dos valores de cada agrupamento.

4.3 Métodos da Análise de Agrupamento

O método utilizado para a análise de agrupamento foi o método hierárquico de Ward para identificação da provável melhor quantidade de grupos e, com isso, foi utilizada essa quantidade para aplicação do método K-*means* no *software* utilizado.

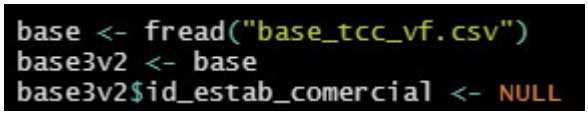
A base gerada da *query* foi extraída no formato de arquivo CSV para ser utilizado no *software* de análises estatísticas RStudio. A base foi importada para o RStudio utilizando o pacote "data.table" que foi criado para manipulação mais rápida de bases de dados grandes, como mostra a figura 13.

Figura 13. Código para instalação do pacote R "data.table"

Fonte: Própria autora

Na figura 14 é demonstrado que foi criada a tabela "base" com o arquivo CSV da tabela importado, na área "Console" do *software*. A terceira linha de código da figura 14 foi utilizada para criação de uma segunda versão da tabela "base", sem o campo "id_estab_comercial", que será utilizado ao fim das análises apenas para identificação de qual grupo cada cliente se enquadra.

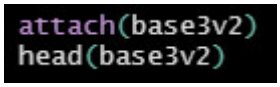
Figura 14. Código de definição da variável "base" com a tabela de dados dos clientes



Fonte: Própria autora

Como se pode ver na figura 15, foi usada a função "attach" para tornar a variável acessível ao digitar seu nome e a função "head" para visualização das primeiras linhas da tabela "base3v2", para verificação de possíveis erros na criação da mesma.

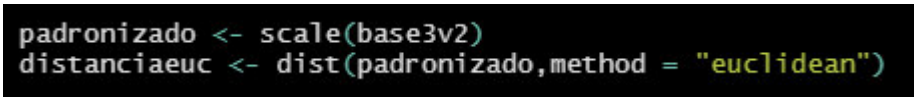
Figura 15. Código das funções attach() e head()



Fonte: Própria autora

A função “scale” foi utilizada com o intuito de padronizar a matriz da tabela criada anteriormente ao subtrair os valores de sua média e dividindo pelo desvio padrão. Para o cálculo das distâncias euclidianas necessárias para o método de Ward, foi utilizada a função “dist” na matriz “padronizado”, utilizando o método Euclidiano, como mostra a figura 16.

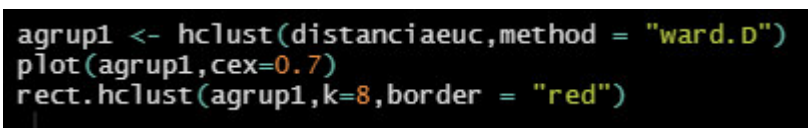
Figura 16. Código de padronização da matriz tabela e cálculo das distâncias euclidianas



Fonte: Própria autora

Após o cálculo das distâncias euclidianas, usou-se a matriz “distanciaeuc” na função “hclust” com o método de Ward, definido anteriormente, para criação da variável “agrup1”. Com isso, a função “plot” foi utilizada para gerar o dendograma criado pelo método de Ward. A função “rect.hclust” foi usada para marcação dos agrupamentos escolhidos no passo anterior, como mostra a figura 17.

Figura 17. Código de agrupamento com o método de Ward e definição da quantidade de agrupamentos



Fonte: Própria autora

Na figura 18 é demonstrado que, depois da definição da quantidade de agrupamentos, foi utilizada a função “kmeans” na matriz “padronizado” para gerar a variável “agrup2”. Este foi utilizado com a função “$cluster” para a definição dos grupos.

Por fim, a função “boxplot” foi utilizada na geração de todos os gráficos de agrupamento de cada variável, média de acessos ao site e ao aplicativo, média de faturamento e vendas e tempo

de empresa do cliente, por ser bom método que indica a localização, a dispersão, a mediana, a assimetria e os outliers dos dados.

Figura 18. Código de utilização do método de K-*means* para separação dos agrupamentos e plotagem dos gráficos

```
agrup2 <- kmeans(padronizado,8)
grupos <- agrup2$cluster
boxplot(md_acessos_site~grupos)
```

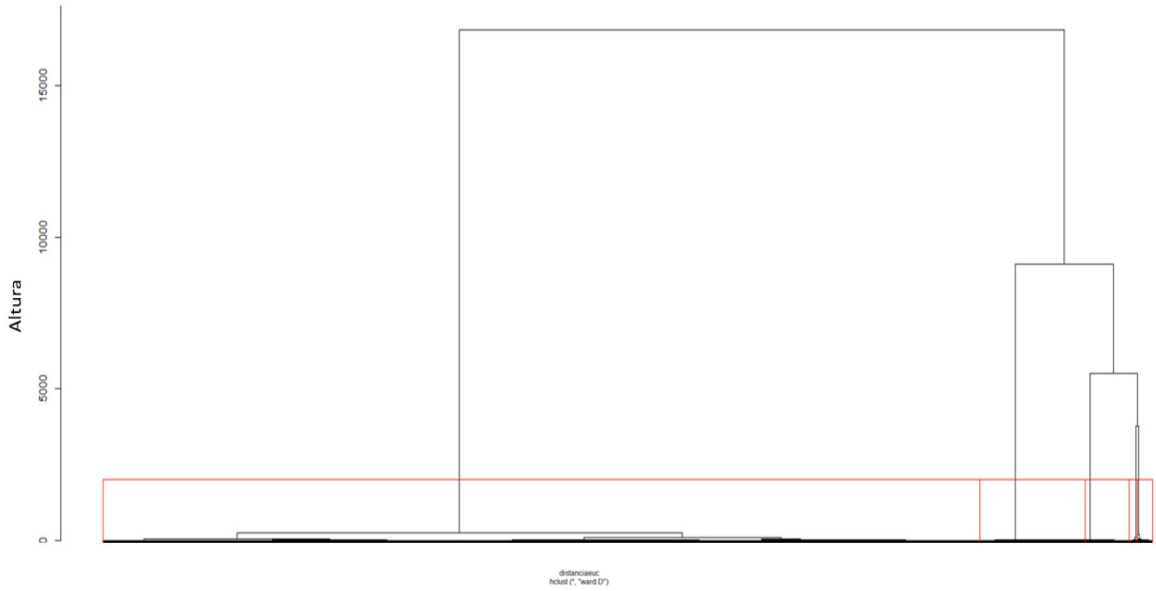
Fonte: Própria autora

5 RESULTADOS E DISCUSSÃO

5.1 Resultados das análises K-*means*

Com a utilização do método de Ward, foi obtido o dendograma da figura 19. No dendograma é possível ver a agrupamento da base em 5 grupos, que estão demarcados pelos retângulos vermelhos para melhor visualização.

Figura 19. Dendograma do agrupamento pelo Método de Ward



Fonte: Própria autora

Na tabela 1, está descrito o tamanho de cada agrupamento. O tamanho dos agrupamentos será relevante apenas para a empresa em estudo entender o esforço que será feito em cada ação, para cada grupo de clientes.

Tabela 1. Descrição da quantidade de estabelecimentos comerciais em cada agrupamento

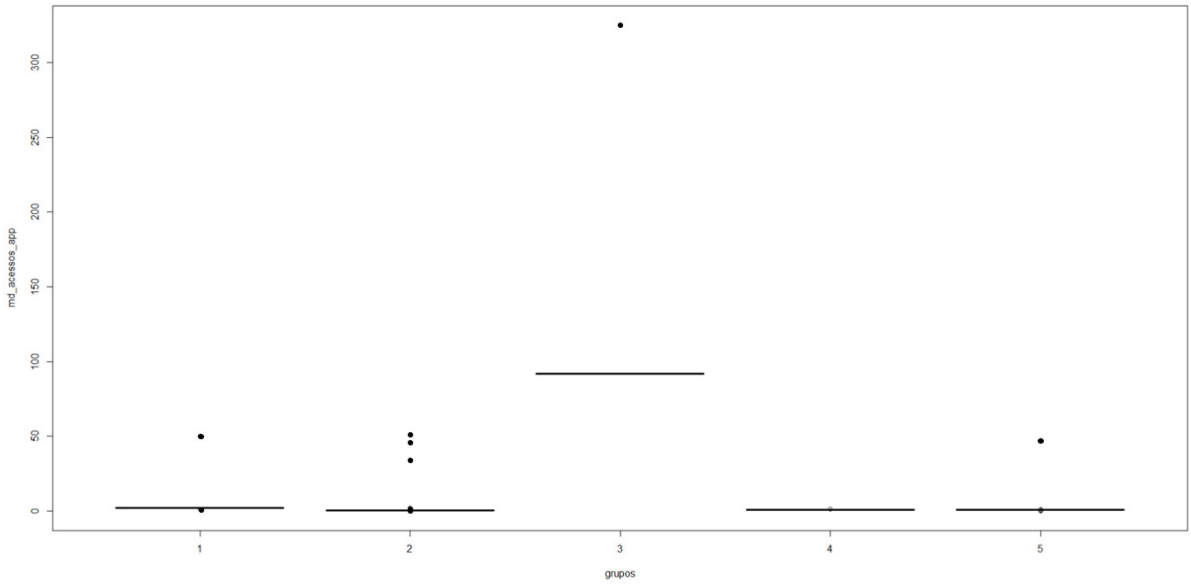
Agrupamento	**Quantidade de Estabelecimentos Comerciais**
Grupo 1	1.060
Grupo 2	370
Grupo 3	194
Grupo 4	2.500
Grupo 5	20.876

Fonte: Própria autora

Com a definição dos 5 grupos da base analisada, a utilização do método de K-*means* com essa quantidade de agrupamento resultou nos gráficos chamados *boxplot* a seguir. Este tipo de gráfico foi escolhido pela sua visualização das medianas, os traços, e dos *outliers* (os círculos) dos dados dos agrupamentos.

No *boxplot* da figura 20, ao analisar a variável de média de acessos ao aplicativo de gestão de negócios da empresa, é possível identificar que o grupo 3 tem uma média de acessos maior do que os outros grupos, com cerca de 92 acessos por mês. Também é possível ver que o grupo 3 tem um *outlier* com mais de 300 acessos. Já os outros grupos têm uma média de quase zero acessos, sendo que o grupo 2 tem mais *outliers*, em torno de 50 acessos por mês.

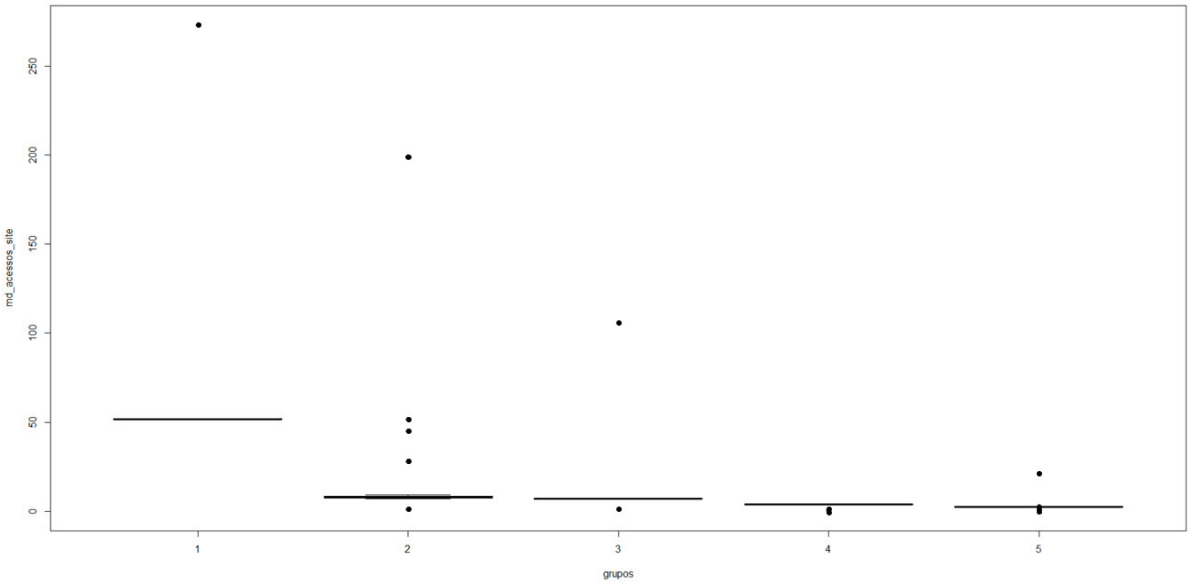
Figura 20. Boxplot dos agrupamentos da média de acessos dos clientes no aplicativo de gestão da empresa



Fonte: Própria autora

Na figura 21, o *boxplot* da variável de média de acessos ao site da empresa mostrou que o grupo 1 tem a maior média, com cerca de 52 acessos por mês, enquanto os outros grupos têm uma média de 5 acessos por mês. Comparando com os acessos ao aplicativo, que a maior parte dos grupos tinha uma média de um acesso por mês, é possível notar que a aderência dos clientes é maior ao site da empresa do que ao aplicativo.

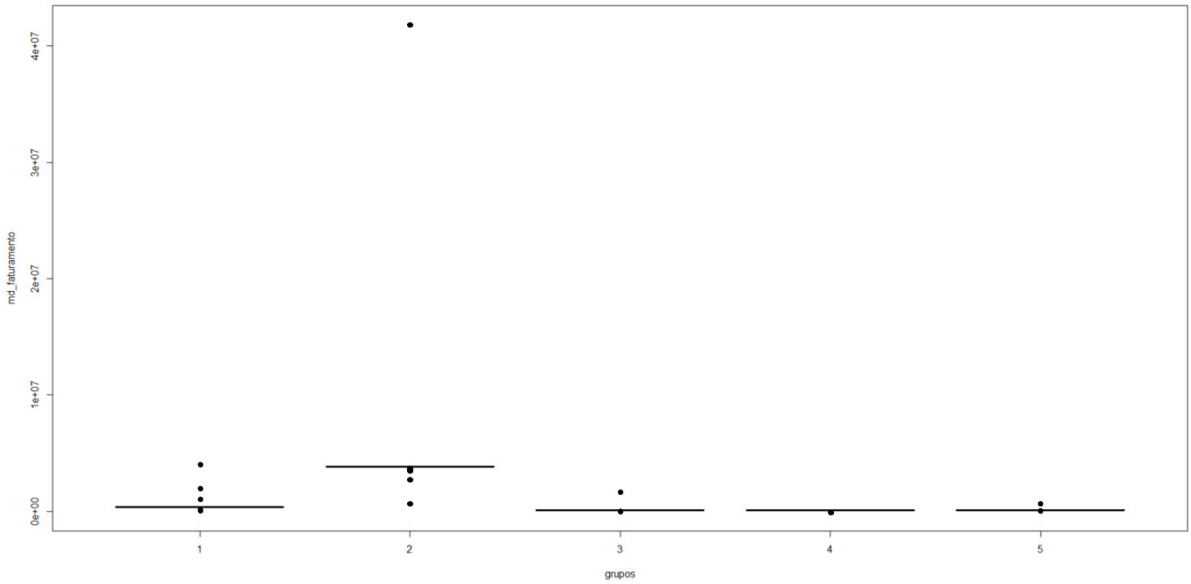
Figura 21. Boxplot dos agrupamentos da média de acessos dos clientes no site de gestão da empresa



Fonte: Própria autora

Os grupos criados na variável de faturamento podem ser visualizados na Figura 22. O grupo 2 tem uma média de R$ 3,9 milhões de faturamento por mês, sendo o grupo com a maior média e com o maior *outlier* também, com cerca de R$ 42 milhões de faturamento. O maior *outlier* do grupo 1 ultrapassa por pouco a média de faturamento do grupo 2. Já os grupos 3, 4 e 5 tem uma média de R$ 177 mil de faturamento.

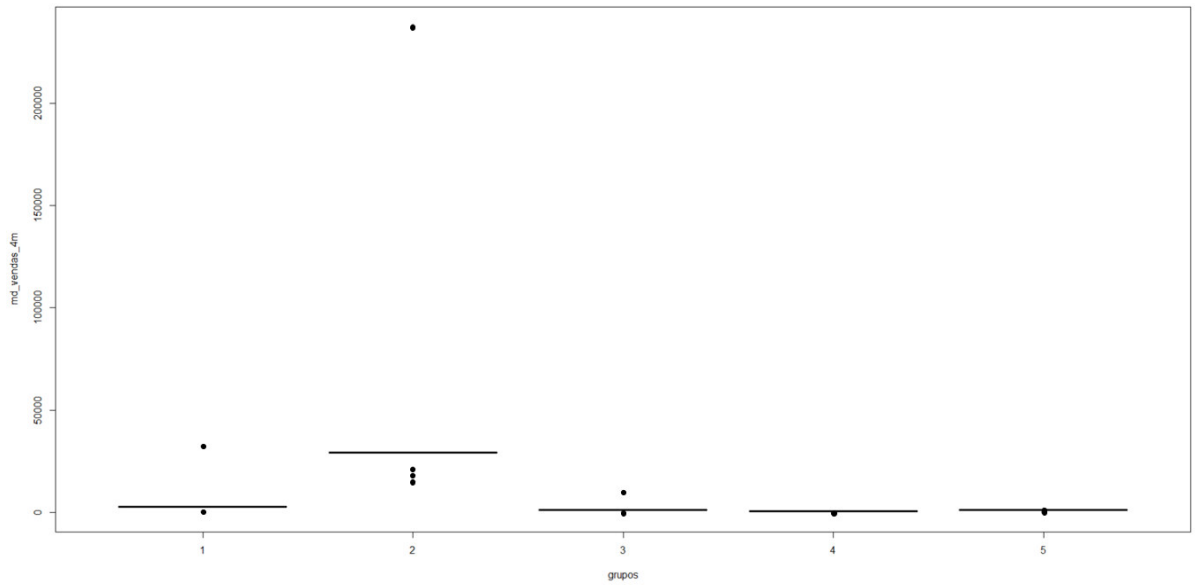
Figura 22. Boxplot dos agrupamentos da média de faturamento dos clientes



Fonte: Própria autora

Seguido do faturamento, é interessante analisar o agrupamento da variável de quantidade de vendas por mês dos clientes. Na figura 23, é possível ver um gráfico bem semelhante com o de faturamento, com o grupo 2 com a maior média de vendas por mês, 29 mil vendas, sugerindo que o *ticket* médio deste grupo é o maior dos grupos. A média de vendas dos outros grupos é de 1.300 vendas por mês.

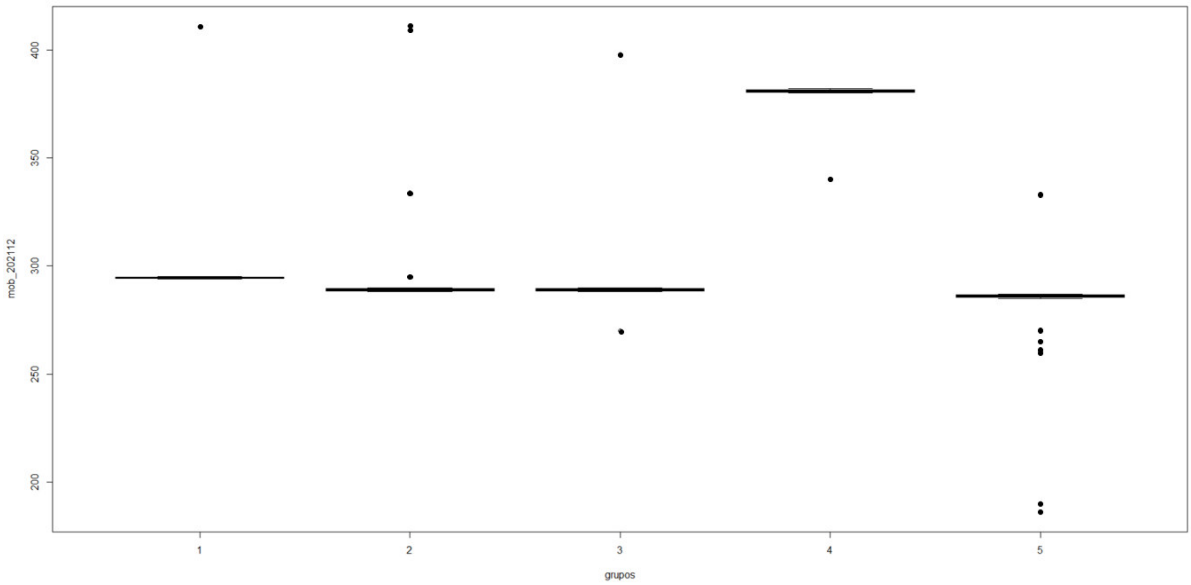
Figura 23. Boxplot da média da quantidade de vendas dos clientes



Fonte: Própria autora

Por fim, a análise de agrupamento da quantidade de meses do cliente na empresa é visualizada no *boxplot* da Figura 24. O grupo 4 é o grupo com maior quantidade de meses na empresa, cerca de 381 meses, ou seja, 31 anos.

Figura 24. Boxplot do tempo de empresa dos clientes

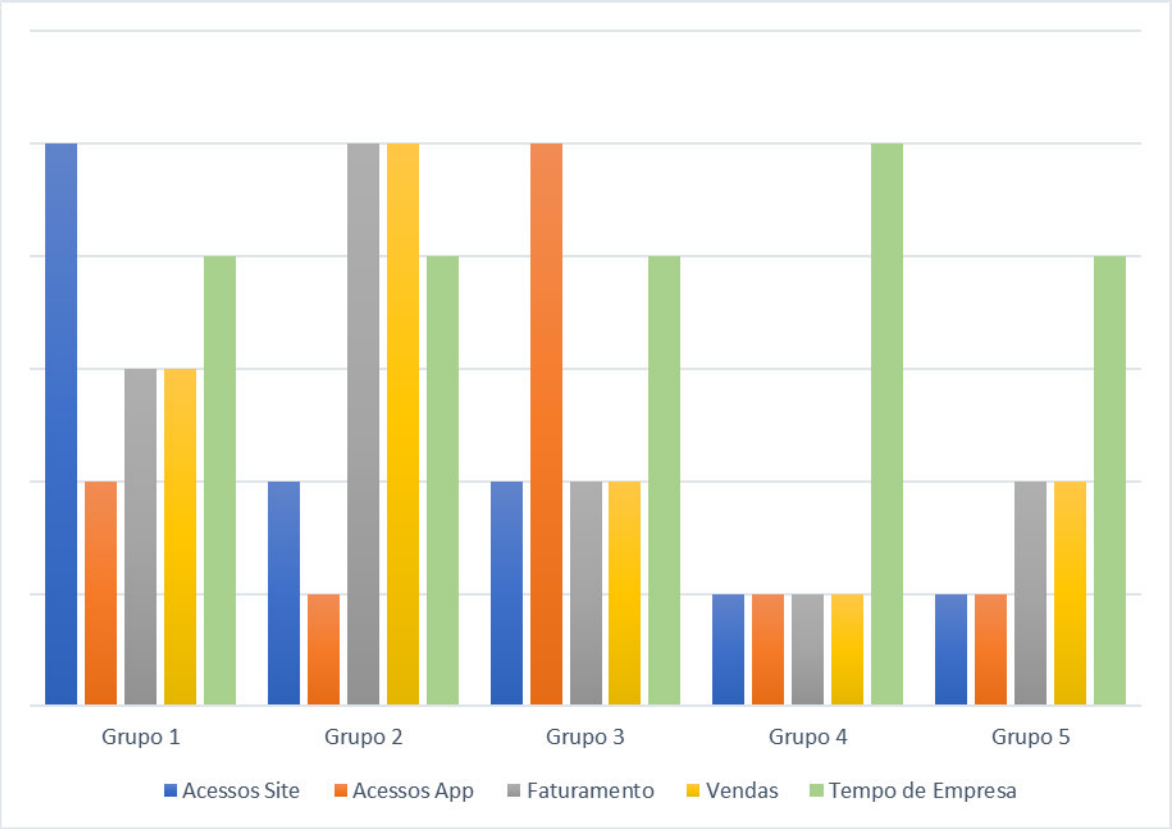


Fonte: Própria autora

5.2 Caracterização dos agrupamentos

Para melhor entendimento dos resultados das análises de cada grupo, na figura 25 a seguir é demonstrado as classificações de cada grupo para cada variável, com relação às médias de cada variável.

Figura 25. Classificação dos agrupamentos de acordo com cada variável de dados



Fonte: Própria autora

A partir da figura 25, foi feita a análise geral de cada agrupamento criado pelo método de K-*means*, para sintetização das características e objetivos das *personas*.

O Grupo 1 foi o grupo mais digitalizado com relação ao site de gestão de vendas da empresa. Sua média de acessos por mês foi muito menor no aplicativo, mas, ainda assim, a média de acessos no aplicativo foi maior do que dos grupos 2, 4 e 5. Pôde-se inferir que é um grupo bem digitalizado e, provavelmente, aderente a campanhas e soluções por meios digitais. Seu *ticket* médio foi o segundo maior dos grupos e teve a segunda maior média de Tempo de Empresa, sugerindo ser uma base de clientes fiéis à empresa.

O Grupo 2 foi o agrupamento com maior faturamento e quantidade de vendas por mês, ou seja, o grupo com maior *ticket* médio. Mas teve a menor média de acessos pelo aplicativo e uma média baixa de acessos pelo site, o que sugere que são clientes que têm sua própria ferramenta de gestão de negócios. Por ser a base de clientes com maior *ticket* médio, é a base de maior potencial para oferecimento de soluções com valores maiores de mensalidade, ou comissão.

O Grupo 3 obteve a maior média de acessos do aplicativo de gestão de vendas da empresa. Como este grupo também obteve a segunda maior média de acessos mensal ao site, é então um grupo com um bom índice de digitalização também, porém, focado no *mobile*. Consequentemente, o Grupo 3, provavelmente, tem boa aderência a soluções por meio de aplicativos. Entretanto, vale ressaltar que este agrupamento obteve o segundo menor ticket médio, logo, soluções oferecidas com alto valor não terão aderência.

O Grupo 4 foi o cluster com a maior média de Tempo de Empresa, porém, obteve os menores índices em todas as outras variáveis, o que sugere que é um grupo de clientes que não concentra suas vendas na empresa em estudo e utilizam outras adquirências simultaneamente. É um bom indicador para focar em campanhas de marketing para aumentar a fidelização do cliente.

Por fim, o Grupo 5 é o agrupamento com a terceira maior média de faturamento e quantidade de vendas por mês. Contudo, tem as menores médias de acesso ao site e ao aplicativo, o que mostra baixa digitalização desse grupo de clientes, possibilitando o direcionamento para soluções que facilitem e otimizem sua gestão de vendas.

Vale ressaltar que, na amostra selecionada da base de clientes ativos, todos são clientes com muito tempo de empresa, que estão afiliados desde o início da mesma.

6 CONCLUSÃO

Com os resultados obtidos da análise de K-*means*, foi possível concluir que o faturamento e a quantidade de vendas são variáveis importantes quanto a agrupamento da base de clientes, para oferecimento de soluções mais condizentes o possível com a realidade de cada estabelecimento comercial em questão. O faturamento foi útil no agrupamento dos estabelecimentos comerciais de acordo com a faixa de preço que é coerente com o investimento que podem fazer nas soluções. Já a quantidade de vendas foi útil no cálculo do *ticket* médio do estabelecimento, ou seja, nos preços que os clientes do estabelecimento estão dispostos a pagar.

O estudo mostrou-se eficaz quanto ao entendimento do perfil de cada cliente estudado ao ser possível entender as *personas* dos cinco agrupamentos criados, com a visualização das características e objetivos dos mesmos.

Uma importante conclusão obtida das análises utilizadas neste estudo é que a amostra é de uma parcela de clientes antigos da empresa, logo, para trabalhos futuros, é importante selecionar uma amostra maior para que a mesma e que consiga abranger todas as *personas* existentes e que, assim, seja possível a criação de campanhas com foco em todos os clientes.

REFERÊNCIAS

ALMEIDA, Marina de Souza. Cartões de crédito: impacto da abertura do mercado de adquirência nas demonstrações financeiras de cielo e redecard. 2013. 19 f. Monografia - Bacharelado em Ciências Contábeis, Universidade de Brasília, Brasília, 2013. Disponível em: https://bdm.unb.br/bitstream/10483/12131/1/2013_MarinadeSouzaAlmeida.pdf. Acesso em: 22 dez. 2021.

APACHE. WHAT IS THE ASF? Disponível em: https://www.apache.org/foundation/. Acesso em: 16 dez. 2021.

AQUINO NETO, Arlindo Fernandes de; ARAUJO, Silvio Roberto Fernandes de. Acelerador do cálculo de distância euclidiana em hardware. 2020. 11 f. TCC (Graduação) - Curso de Ciência da Computação, Universidade Federal Rural do Semi-Árido, Mossoró, 2021. Disponível em: http://repositorio.ufersa.edu.br/handle/prefix/6465. Acesso em: 14 dez. 2021.

AWS. Hue. Disponível em: https://docs.aws.amazon.com/pt_br/emr/latest/ReleaseGuide/emr-hue.html. Acesso em: 15 dez. 2021.

BARROS, Giulia Gonçalves de. Personas da vida real: um framework para criação de Personas em projetos e suas limitações. 2019. Dissertação (Mestrado em Design) - Universidade Federal de Pernambuco, Recife, 2019.

CLOUDERA. Introduction to Hue. Disponível em: https://docs.cloudera.com/documentation/enterprise/6/6.3/topics/hue.html. Acesso em: 16 dez. 2021.

FABIOBMED. Inner Join, Left Join e Right Join. 2013. Disponível em: https://www.fabiobmed.com.br/site/inner-join-left-join-e-right-join/. Acesso em: 27 nov. 2021.

MINGOTI, S. A.; Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada, Editora UFMG, 2005.

Moori, Roberto Giro, Marcondes, Reynaldo Cavalheiro e Ávila, Ricardo Teixeira. A análise de agrupamentos como instrumento de apoio à melhoria da qualidade dos serviços aos clientes. Revista de Administração Contemporânea [online]. 2002, v. 6, n. 1, pp. 63-84. Disponível em:

<https://doi.org/10.1590/S1415-65552002000100005>. Epub 30 Mar 2009. ISSN 1982-7849. Acesso em: 28 nov. 2021.

NIELD, Thomas. Introdução à Linguagem SQL: abordagem prática para iniciantes. São Paulo: Novatec Editora, 2016. 144 p.

NISHI, Luciana; SOUZA, Fernando José de; SANTANA, Paulo Henrique Alves de. Estudo de caso: análise entre banco de dados relacional e não relacional. 2017. 32 f. TCC (Graduação) - Curso de Engenharia de Computação, Centro Unversitário de Anápolis – Unievangélica, Anápolis, 2017. Disponível em: http://repositorio.aee.edu.br/jspui/handle/aee/46. Acesso em: 15 dez. 2021.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

RITTER, Matias do Nascimento et al. Introdução ao software estatístico R. Imbé: Creative Commons, 2019. 110 p. Disponível em: https://www.lume.ufrgs.br/bitstream/handle/10183/205058/001108586.pdf?sequence=1. Acesso em: 5 dez. 2021.

SEIDEL, Enio Júnior et al. Comparação entre o método Ward e o método K-médias no agrupamento de produtores de leite. Ciência e Natura, Santa Maria, v. 1, n. 30, p. 7-15, dez. 2008. Disponível em: http://www2.assis.unesp.br/ffrei/Artigos/Compara%C3%A7%C3%A3o%20entre%20o%20m%C3%A9todo%20Ward%20e%20o%20m%C3%A9todo%20K-m%C3%A9dias%20no%20agrupamento%20de%20produtores%20de%20leite.pdf. Acesso em: 29 nov. 2021.

SILVA, Jéssica Lobo Rodrigues da. Perfil de Clientes de uma Empresa de Entregas. 2019. 58 f. TCC (Graduação) - Curso de Engenharia Física, Universidade de São Paulo, Lorena, 2019. Disponível em: https://sistemas.eel.usp.br/bibliotecas/monografias/2019/MEF19007.pdf. Acesso em: 21 nov. 2021.

VALLI, Márcio. Análise de Cluster. Augusto Guzzo Revista Acadêmica, São Paulo, n. 4, p. 77-87, aug. 2012. ISSN 2316-3852. Disponível em: <http://www.fics.edu.br/index.php/augusto_guzzo/article/view/107>. Acesso em: 29 nov. 2021.