CAUÊ LUNA DA SILVA

CLUSTERIZAÇÃO E OTIMIZAÇÃO DE BASE DE CLIENTES PARA MAXIMIZAÇÃO DA ADESÃO À PROPOSTA DE VALOR DA ORGANIZAÇÃO: ESTUDO DE CASO EM STARTUP DO SETOR DE VAREJO ALIMENTAR

São Paulo
2022

CAUÊ LUNA DA SILVA

CLUSTERIZAÇÃO E OTIMIZAÇÃO DE BASE DE CLIENTES PARA MAXIMIZAÇÃO DA ADESÃO À PROPOSTA DE VALOR DA ORGANIZAÇÃO: ESTUDO DE CASO EM STARTUP DO SETOR DE VAREJO ALIMENTAR

Trabalho de formatura apresentado à Escola Politécnica da Universidade de São Paulo para a obtenção do diploma de Engenheiro de Produção.

São Paulo
2022

CAUÊ LUNA DA SILVA

CLUSTERIZAÇÃO E OTIMIZAÇÃO DE BASE DE CLIENTES PARA MAXIMIZAÇÃO DA ADESÃO À PROPOSTA DE VALOR DA ORGANIZAÇÃO: ESTUDO DE CASO EM STARTUP DO SETOR DE VAREJO ALIMENTAR

Trabalho de Formatura apresentado à Escola Politécnica da Universidade de São Paulo para a obtenção do diploma de Engenheiro de Produção.

Orientador: Prof. Dr. Eduardo de Senzi Zancul

São Paulo
2022

FICHA CATALOGRÁFICA

Da Silva, Cauê Luna.

Clusterização e Otimização de base de clientes para maximização da adesão à proposta de valor da organização: estudo de casa em startup do setor de varejo alimentar / Cauê Luna da Silva - São Paulo, 2022. 148p.

Trabalho de Formatura - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Produção.

1. Varejo Alimentar 2. Startup 3. Perfil de Cliente 4. Proposta de Valor 5. Clusterização 6. K-Means I. Cauê Luna da Silva II. Escola Politécnica da Universidade de São Paulo III. Título

À minha família e aos meus amigos que
me acompanharam nesta trajetória.

AGRADECIMENTOS

À minha família, que proporcionou todo o apoio e a estrutura necessários para meu crescimento pessoal e profissional ao longo de minha trajetória, sem os quais seria incapaz de ter sequer iniciado esta graduação. Agradeço por todos os valores e ensinamentos agregados à minha vida. Agradeço também por terem acreditado e investido na minha formação, além de terem provido todo o apoio necessário para esta jornada.

Aos amigos que fiz dentro e fora da Escola Politécnica da USP, sem o apoio dos quais seria ainda mais árduo o processo que passei ao longo destes anos de formação. Agradeço também ao Centro Acadêmico de Engenharia de Produção e à Poli Júnior, instituições que foram essenciais para minha formação como engenheiro.

Aos profissionais da empresa para a qual foi desenvolvido este projeto, que me guiaram ao longo dos últimos meses para assegurar que valor entregue em todas as etapas fosse convertido em impacto significativo para a empresa e para os nossos clientes. Agradeço por todo o apoio, por todos os ensinamentos e por toda a confiança depositada em meu trabalho.

A todos os docentes da Escola Politécnica da USP que lutam pelo desenvolvimento desta renomada universidade, pela defesa da educação e pela defesa da ciência. Agradeço a todos aqueles que buscam atualizar e evoluir os métodos de ensino desta universidade, evitando a obsolescência das práticas pedagógicas nela aplicadas e assegurando o reconhecimento internacional da instituição.

RESUMO

O mercado de varejo alimentar brasileiro passa por um cenário de elevada competitividade e de digitalização de operações antes feitas de modo analógico. Neste contexto, indústrias, distribuidores e atacadistas buscam cada vez mais compreender seu perfil de cliente para desenvolver seus produtos e serviços de modo que estes sejam capazes de agregar elevado valor a seus clientes. A empresa para qual foi realizado este estudo de caso buscou aumentar as barreiras de entrada para concorrentes e reduzir o poder de barganha de seus consumidores em seu mercado-alvo no intuito de consolidar suas operações de forma lucrativa e que permita uma operação sólida no longo prazo.

Tendo como base um modelo de negócios voltado para o abastecimento de produtos de mercearia na cidade de São Paulo e tomando as características específicas da operação da empresa em consideração foi inicialmente desenvolvido um modelo capaz de segmentar os clientes da empresa em grupos distintos de acordo com o perfil de compra de cada cliente. Foram também analisados os potenciais de compra máximo de cada perfil de cliente e as características de maior correlação com cada um destes grupos para um aprofundamento no conhecimento acerca de cada um destes.

Tendo conhecidos os diferentes perfis de cliente existentes neste mercado, foi definida a composição ideal da carteira de clientes da empresa no longo prazo para a garantia de que essa carteira esteja adequada para a proposta de valor oferecida pela empresa. Para auxiliar a empresa a migrar da composição atual de base de clientes para esta composição ideal, foram analisadas as transições dos clientes no passado a fim de entender os fatores que devem ser impulsionados para aumentar a incidência de transições positivas e os fatores que devem ser combatidos para reduzir a incidência de transições negativas.

Todas as iniciativas realizadas neste projeto convergem, por fim, para a criação de planos de ação recomendados pelo autor à empresa cliente para a consolidação de operações mais rentáveis. Estes planos de ação estão em implementação na organização e os resultados estão sendo acompanhados de maneira próxima pelos principais *stakeholders* do projeto.

Palavras-chave: Varejo Alimentar, Startup, Perfil de Cliente, Proposta de Valor, Clusterização, *K-Means*.

ABSTRACT

The Brazilian food retail market is passing through a scenario of high competitiveness and digitalization of operations that were previously made in analogue forms. In this context, industries, distributors and cash and carries seek to understand their customer profile to develop their products and services in a way that they become capable of generating high added value to those customers. The company for which this case study was developed seeked to increase the barriers to new entrants and to reduce the bargaining power of buyers on its market in order to consolidate its operations in a profitable way that enables a solid operation in the long term.

Having a business model focused on the supply of grocery products in the city of São Paulo and taking the specific characteristics of the company in consideration, a machine learning model was initially developed to clusterize the company's customers in distinct groups in accordance with each customer's purchasing profile. The customer's maximum purchase potential and its characteristics with higher correlation with its clusterization were also analyzed in order to provide a deepening in each of those customers.

Having analyzed each of the customer profiles that exist on this market, the company's ideal composition of the customer portfolio in the long term was also defined to guarantee that its customers portfolio become proper to the value proposition offered by the company. To back up the company's migration from the current customer portfolio composition to the ideal one, the transitions of customers between clusters in the past were also analyzed in order to understand which factors should be boosted to increase the positive transitions and which factors should be fought against to decrease the negative transitions.

In the end, all the initiatives accomplished in this project converge to the development of action plans recommended by the author to the client company to consolidate more profitable operations. These action plans are currently in implementation at the organization and the results are being followed closely by the main stakeholders of the project.

Keywords: Food Retail, Startup, Customer Profile, Value Proposition, Clusterization, K-Means

LISTA DE ILUSTRAÇÕES

LISTA DE GRÁFICOS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

OKRS	Objetivos e Resultados Chave
SQL	Linguagem de Consulta Estruturada
KPI	Indicador-Chave de Desempenho
SKU	Unidade de Manutenção de Estoque
PIB	Produto Interno Bruto
SEBRAE	Serviço Brasileiro de Apoio às Micro e Pequenas Empresas
ABRAS	Associação Brasileira de Supermercados
FLV	Frutas, Legumes e Verduras
MVV	Missão, Visão e Valores
CRF	Ciclo de Conversa, Reconhecimento e *Feedback*
SCQ	Situação, Complicação e Questão
MECE	Mutuamente Excludente e Coletivamente Exaustivo
DRE	Demonstração de Resultados
IDE	Ambiente de Desenvolvimento Integrado
API	Interface de Programação de Aplicativos
NPS	Escala de Promoção da Rede
DBSCAN	Clusterização Espacial Baseada em Densidade de Aplicações com Ruído
K-NN	K Vizinhos Mais Próximos
E-Commerce	Comércio Eletrônico
ERP	Sistema de Planejamento de Recursos Empresariais
WMS	Sistema de Gerenciamento de Estoques
CRM	Gestão de Relacionamento com o Cliente
TI	Tecnologia da Informação

SUMÁRIO

1 INTRODUÇÃO

Na introdução serão abordados os aspectos inerentes à análise desenvolvida nos capítulos subsequentes. Inicialmente, será apresentado um contexto do trabalho que será contemplado pelas caracterizações do mercado analisado, do perfil-alvo de clientes, da empresa onde o estudo foi desenvolvido, da proposta de valor da empresa e dos produtos e serviços oferecidos aos clientes. De maneira complementar, serão apresentados a motivação, o problema e os objetivos que sustentam a necessidade do desenvolvimento do presente trabalho para a empresa-alvo do estudo. E, por fim, será apresentada a estrutura do trabalho para melhor detalhar o modo como o problema em questão será atacado ao longo dos capítulos seguintes no intuito de melhor orientar o leitor ao longo do material elaborado.

1.1 Contexto do Trabalho

O desenvolvimento das análises, das modelagens, e das iniciativas expostas nos capítulos seguintes foram realizadas em consonância com a rotina do autor durante suas atividades na empresa cliente do projeto enquanto Analista de Dados. Desse modo, faz-se necessário apresentar ao leitor um contexto inicial acerca dos principais aspectos que permeiam o contexto em que o projeto foi desenvolvido para que seja possível que este compreenda adequadamente as motivações e os objetivos, bem como o problema que serão descritos nos tópicos a seguir ainda dentro desta introdução. Desse modo, serão apresentados - para cada um dos aspectos inerentes ao contexto mencionado - os principais dados e percepções que resumem o modelo de negócio da empresa para a qual o projeto foi desenvolvido.

1.1.1 Mercado Analisado

A empresa estudada atua no ramo de varejo alimentar no estado de São Paulo. Para caracterizar este mercado a fim de compreender adequadamente o ramo de atuação da empresa, é necessário compreender a relevância do varejo no PIB Brasileiro, assim como a relevância do varejo alimentar dentro do varejo geral e a relevância do varejo de proximidade - categoria-alvo da empresa - dentro do varejo alimentar. De modo complementar, será

discorrido sobre a relevância do estado de São Paulo no varejo alimentar brasileiro e sobre as tendências deste segmento de mercado no Brasil para os próximos anos. Com estes pontos, espera-se que o leitor familiarize-se inicialmente com o mercado em que serão pautados os tópicos seguintes.

Para analisar a representatividade do Varejo no PIB Brasileiro, foram utilizados os dados oficiais do Sistema de Contas Nacionais Trimestrais (IBGE, 2021) e sua respectiva análise elaborada no estudo "O Papel do Varejo na Economia Brasileira" (SBVC, 2021). O principal indexador utilizado para analisar o varejo frente à economia brasileira é o dado de Consumo das Famílias, que é apurado trimestralmente no estudo do Instituto Brasileiro de Geografia e Estatística supracitado. Em 2020, o Consumo das Famílias representou R$4,67 trilhões (Tabela 1), o que representa aproximadamente 61% do Produto Interno Bruto do ano.

Tabela 1: Classes de atividade no valor adicionado a preços básicos e componentes do PIB pela ótica da despesa

Classes de atividade no valor adicionado a preços básicos e componentes do PIB pela ótica da despesa (R$ milhões)						
Especificação	2019	2020.I	2020.II	2020.III	2020.IV	2020
Agropecuária	R$ 326.040,00	R$ 124.866,00	R$ 127.239,00	R$ 105.459,00	R$ 82.275,00	R$ 439.838,00
Indústria	R$ 1.363.547,00	R$ 313.521,00	R$ 302.755,00	R$ 354.045,00	R$ 344.234,00	R$ 1.314.555,00
Serviços	R$ 4.680.170,00	R$ 1.143.671,00	R$ 1.103.492,00	R$ 1.168.093,00	R$ 127.114,00	R$ 4.686.370,00
Valor Adicionado a Preços Básicos	R$ 6.369.757,00	R$ 1.582.058,00	R$ 1.533.485,00	R$ 1.627.597,00	R$ 1.697.623,00	R$ 6.440.763,00
Impostos sobre produtos	R$ 1.037.267,00	R$ 261.805,00	R$ 175.275,00	R$ 264.138,00	R$ 305.877,00	R$ 1.007.095,00
PIB a Preços de Mercado	R$ 7.407.024,00	R$ 1.843.863,00	R$ 1.708.760,00	R$ 1.891.735,00	R$ 2.003.500,00	R$ 7.447.858,00
Despesa de Consumo das Famílias	4797,18	R$ 1.184.872,00	R$ 1.038.340,00	R$ 1.167.913,00	R$ 1.279.785,00	R$ 4.670.910,00
Despesa de Consumo do Governo	R$ 1.487.164,00	R$ 349.885,00	R$ 377.507,00	R$ 371.233,00	R$ 427.658,00	R$ 1.526.283,00
Formação Bruta de Capital Fixo	R$ 1.134.200,00	R$ 293.311,00	R$ 257.463,00	R$ 306.322,00	R$ 366.638,00	R$ 1.223.733,00
Exportações de Bens e Serviços	R$ 1.044.787,00	R$ 260.691,00	R$ 324.086,00	R$ 336.965,00	R$ 334.775,00	R$ 1.256.517,00
Importações de Bens e Serviços (-)	R$ 1.062.950,00	R$ 280.418,00	R$ 263.764,00	R$ 272.610,00	R$ 336.393,00	R$ 1.153.185,00
Variação de Estoque	R$ 6.705,00	R$ 35.522,00	-R$ 24.873,00	R$ 18.087,00	-R$ 68.963,00	-R$ 76.401,00

Fonte: Instituto Brasileiro de Geografia e Estatística, 2020

Ao analisarmos de maneira mais profunda o setor de varejo brasileiro, podemos detalhar a representatividade do varejo alimentar dentro deste setor e também dentro do PIB brasileiro de forma geral. Com base no Ranking ABRAS (Associação Brasileira de Supermercados, 2022), o Varejo Alimentar alcançou um faturamento de R$554 bilhões (Tabela 2), o que representa 7,5% do PIB nacional e aproximadamente 11,86% do faturamento do setor de varejo - quando considerado em comparação com os dados do

Sistema Nacional de Contas Trimestrais. Essa pesquisa retrata a relevância do mercado analisado.

Tabela 2: Faturamento anual do varejo alimentar brasileiro

Ano	Faturamento Anual (R$ bilhões)
2011	224,3
2012	242,9
2013	272,2
2014	294,9
2015	316,2
2016	338,7
2017	353,2
2018	355,7
2019	378,3
2020	554,0

Fonte: Ranking ABRAS 2021

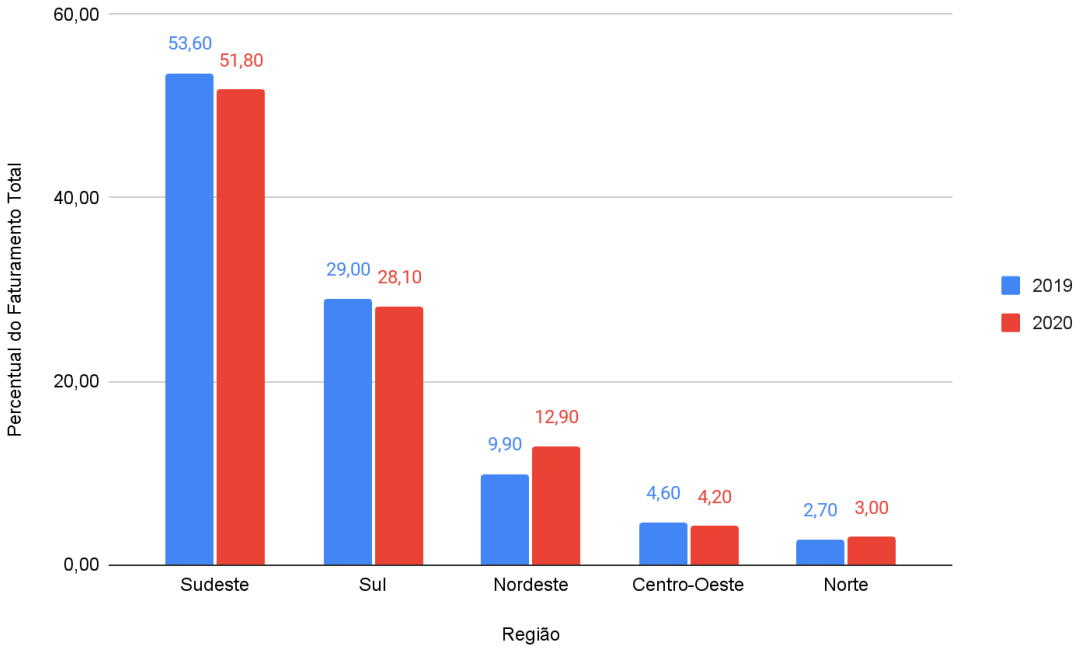
O Varejo Alimentar brasileiro, por sua vez, pode ser decomposto em quatro micro categorias em uma partição que é essencial para a devida compreensão do mercado de atuação da empresa-alvo do estudo. Estas categorias são: Supermercados e Hipermercados, *Cash & Carry* (Atacarejos), Tradicional e Conveniência. Em nosso contexto, o público-alvo do projeto desenvolvido é centrado no Varejo Tradicional, que também pode ser referenciado como Varejo de Proximidade.

O Varejo de Proximidade é composto por mercados que comercializam produtos alimentares industrializados, ou não, com até quatro caixas registradores ou faturamento de até R$3,6 milhões por ano, segundo o presidente do Sebrae, Luiz Barretto Filho. Este grupo corresponde a 35% das vendas do setor supermercadista nacional e é composto por cerca de 416 mil estabelecimentos (SEBRAE, 2015). Essa participação confere ao setor um faturamento anual aproximado de R$190 bilhões sendo, portanto, umas das parcelas mais relevantes da economia brasileira tanto em faturamento quanto em número de lojas. Vale

também ressaltar que o Varejo de Proximidade possui diversos estabelecimentos informais em território nacional, o que pode fazer com que o número real de mercados que se encaixam nesta categoria seja ainda maior do que o valor indicado pelos números oficiais do SEBRAE.

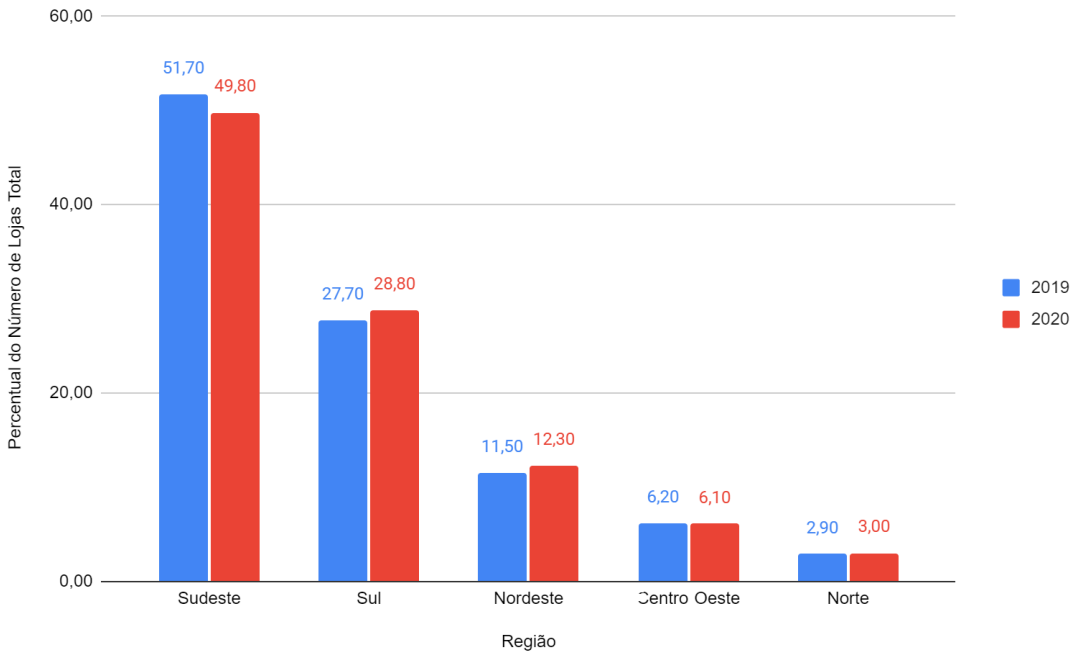
Ao analisarmos o setor de Varejo Alimentar brasileiro podemos realizar recortes regionais e estaduais a fim de melhor compreendermos a distribuição da representatividade deste setor no território nacional, bem como melhor caracterizar o mercado em que Varejo ABC está inserida. No Ranking ABRAS (Associação Brasileira de Supermercados, 2022) é possível identificarmos que a região sudeste é a mais expressiva no cenário nacional, sendo responsável por 51,8% do faturamento (Gráfico 1) e por 49,8% das lojas do Varejo Alimentar Brasileiro (Gráfico 2).

Gráfico 1: Distribuição do faturamento do varejo alimentar brasileiro por região



Fonte: Ranking ABRAS 2021

Gráfico 2: Distribuição do número de lojas do varejo alimentar brasileiro por região



Fonte: Ranking ABRAS 2021

A partir da mesma pesquisa que baseia os gráficos acima, é possível extrair um recorte acerca da representatividade de cada estado brasileiro no faturamento do setor tomando como base os dados do ano de 2020. O estado de São Paulo é o maior representante do setor, com 34,6% do faturamento total levantado na pesquisa (Tabela 3), ficando muito à frente do segundo colocado - Minas Gerais - que possui 11,6% do faturamento.

Tabela 3: Distribuição do faturamento do varejo alimentar brasileiro por estado

Class. 2020	Estado	Faturamento	%
1	SP	R$ 125.527.002.665,00	34,6
2	MG	R$ 41.941.502.576,00	11,6
3	RS	R$ 38.560.321.621,00	10,6
4	PR	R$ 36.040.397.084,00	9,9
5	SC	R$ 27.279.108.786,00	7,5
6	RJ	R$ 19.215.863.231,00	5,3
7	MA	R$ 13.335.936.638,00	3,7
8	CE	R$ 8.294.504.591,00	2,3
9	BA	R$ 7.370.386.789,00	2,0
10	PA	R$ 6.638.207.298,00	1,8
11	RN	R$ 4.689.657.775,00	1,3
12	DF	R$ 4.261.753.423,00	1,2
13	MS	R$ 4.193.451.618,00	1,2
14	PE	R$ 3.930.012.436,00	1,1
15	GO	R$ 3.886.334.693,00	1,1
16	PI	R$ 3.830.907.528,00	1,1
17	PB	R$ 2.908.164.230,00	0,8
18	MT	R$ 2.842.672.790,00	0,8
19	AL	R$ 1.622.708.578,00	0,4
20	AC	R$ 1.547.397.223,00	0,4
21	ES	R$ 1.314.436.906,00	0,4
22	AP	R$ 900.408.452,00	0,2
23	SE	R$ 898.403.995,00	0,2
24	TO	R$ 796.711.820,00	0,2
25	RO	R$ 506.109.966,00	0,1
26	AM	R$ 310.093.308,00	0,1
27	RR	-	0,0
Total		R$ 362.642.456.021,00	100

Fonte: Ranking ABRAS 2021

Feitas as contextualizações acima, é possível compreendermos melhor o mercado em que a empresa cliente do projeto desenvolvido neste trabalho está inserida. Sendo possível contextualizar tanto o mercado de Varejo Tradicional - partindo das análises do Varejo Alimentar e do Varejo Geral - quanto a representatividade do estado de São Paulo, onde estão concentradas as operações atuais da empresa, é definida a pertinência deste contexto de

mercado de atuação para o foco inicial. Cabe agora detalhar com maior profundidade o perfil dos clientes que compõem o Varejo Tradicional.

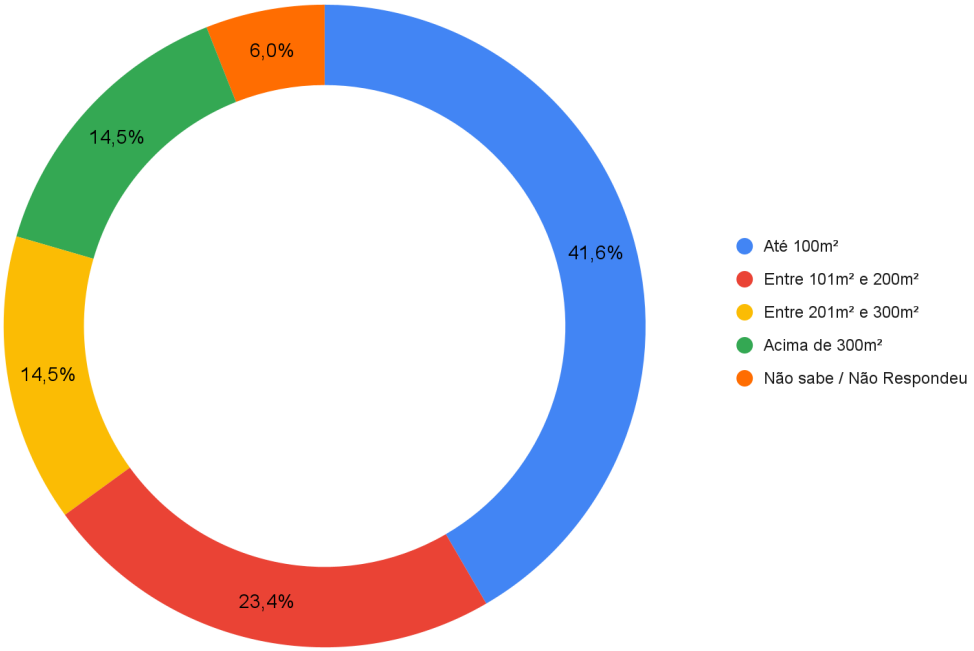
1.1.2 Perfil de Clientes

O projeto desenvolvido precisa atender de maneira personalizada o nicho de mercado focado pela empresa para o qual ele foi desenvolvido. Desse modo, será introduzido nesta seção o perfil de cliente que será focado para as construções subsequentes, lembrando sempre que este cliente está inserido dentro do contexto do Varejo Tradicional definido no tópico anterior. Nesta seção se discorrerá brevemente sobre as principais dores enfrentadas por este perfil de cliente em sua rotina.

Durante a caracterização do perfil dos clientes deste segmento, serão analisadas diversas características fundamentais para a divisão destes em grupos de perfil semelhante que tendem a ter também semelhantes comportamentos e perfis de compra. Essas características, assim como muitas outras a serem compiladas no levantamento de dados realizado a posteriori, serão utilizadas ao longo do projeto para modelar a adesão destes clientes à proposta de valor da empresa e para definir as iniciativas priorizadas para o melhor aproveitamento deste potencial.

O Varejo Tradicional compreende, em geral, lojas de tamanho bastante reduzido. No levantamento Pesquisa de Minimercados no Brasil (SEBRAE, 2015), podemos observar que 41,6% dos varejistas desta categoria possuem lojas de até 100 m² e 23,4% destes possuem lojas entre 100 m² e 200 m² (Gráfico 3). Esse dado gera diversas implicações inerentes a atributos como a capacidade de armazenamento de produtos, ao giro do estoque e a capacidade de atendimento que esses lojistas possuem.

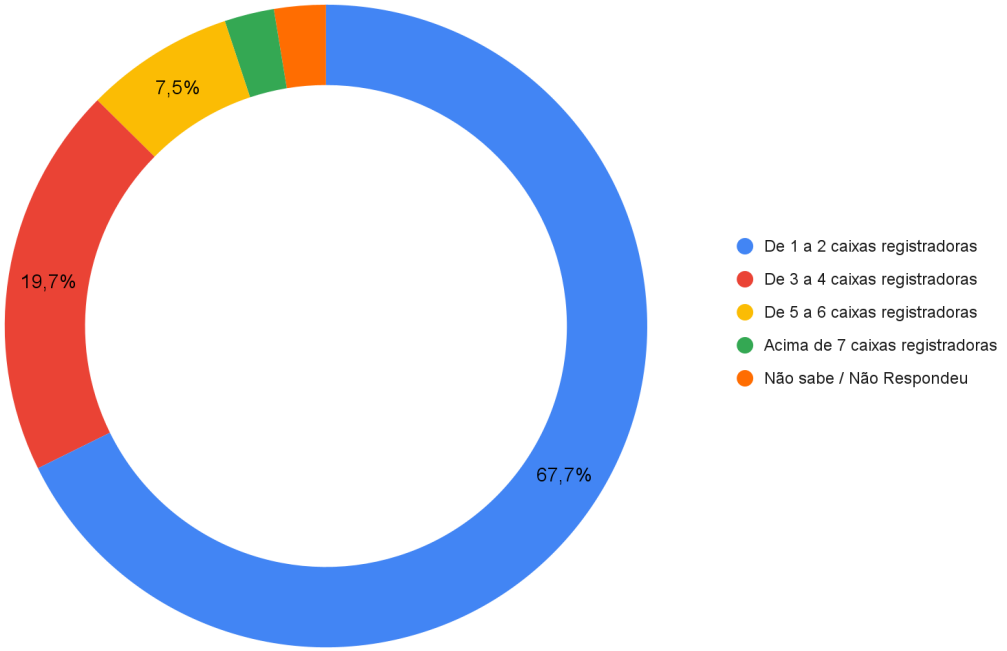
Gráfico 3: Distribuição dos varejistas tradicionais por tamanho de loja



Fonte: Pesquisa de Minimercados no Brasil, 2015

Além disso, tomando por base o número de caixas registradoras (checkouts) que cada loja possui - que é um dos parâmetros mais usuais no ramo do varejo tradicional para analisar o porte de lojas - podemos identificar que dentre os potenciais clientes analisados, 67,7% possuem apenas entre 1 e 2 caixas registradoras em suas lojas (Gráfico 4).

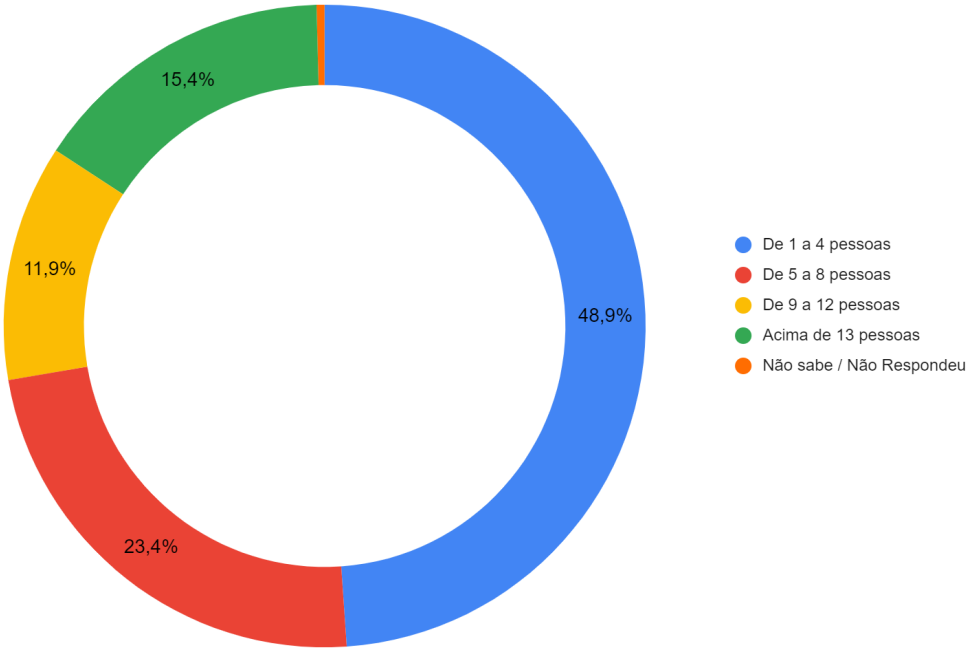
Gráfico 4: Distribuição dos varejistas tradicionais por número de caixas



Fonte: Pesquisa de Minimercados no Brasil, 2015

Outro aspecto bastante relevante para a gestão das lojas e consequentemente para a performance dos lojistas tanto em compras e vendas quanto em outros aspectos organizacionais é o número de funcionários que a loja possui. Como podemos observar no Gráfico 5, é notório que o perfil de varejistas tradicionais está predominantemente concentrado em lojas com um corpo de funcionários bastante reduzido, sendo que 48,9% das lojas possuem 4 ou menos funcionários. Inclusive, uma característica muito marcante deste perfil de estabelecimento no Brasil é o fato de muitos deles possuírem uma gestão familiar, em que a loja é gerida predominantemente por membros de uma mesma família e ocorrendo ainda diversos casos em que a loja passa por diversas gerações da família.

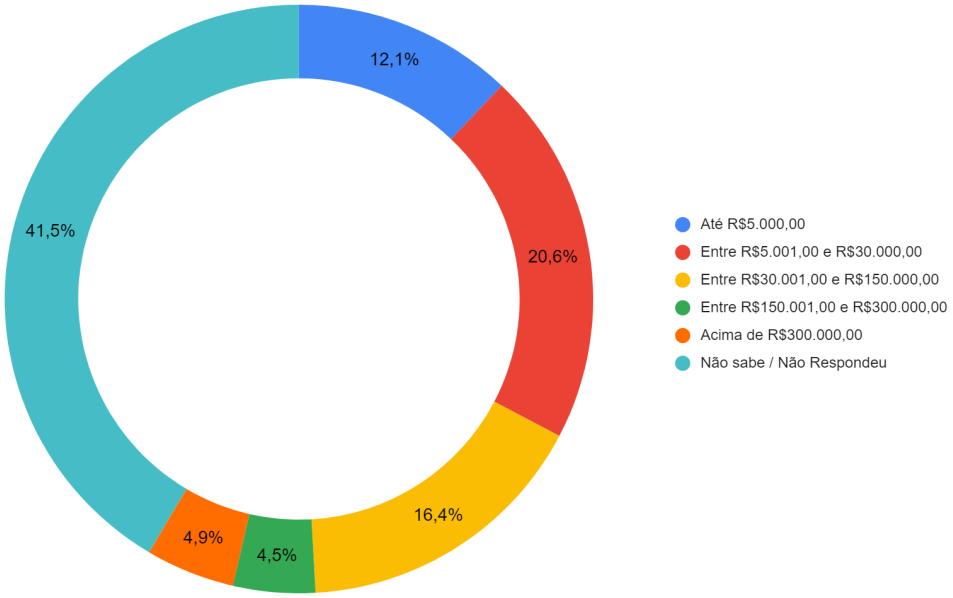
Gráfico 5: Distribuição dos varejistas tradicionais por número de funcionários



Fonte: Pesquisa de Minimercados no Brasil, 2015

Com base nos dados anteriormente levantados acerca do faturamento e do número de lojas neste setor e dos dados levantados neste tópico acerca do porte das lojas que compõem esse setor, se faz previsível o fato de que, embora o setor possua uma enorme representatividade na economia nacional em termos de faturamento, este é muito pulverizado entre as diversas lojas existentes no Brasil. Podemos ver que o faturamento mensal de uma parcela bastante relevante dos lojistas deste segmento sequer ultrapassa R$100 mil (Gráfico 6). Este fator intensifica a necessidade de estratégias muito bem elaboradas para o aproveitamento deste potencial de compra existente no mercado sem incorrer em custos operacionais extremamente elevados por conta do atendimento pulverizado de clientes.

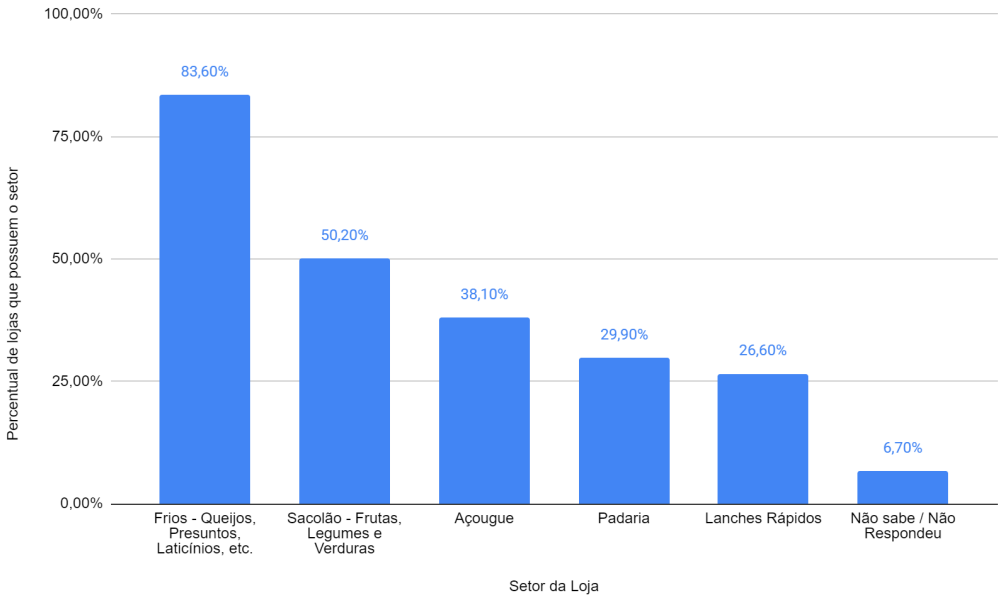
Gráfico 6: Distribuição dos varejistas tradicionais por faturamento



Fonte: Pesquisa de Minimercados no Brasil, 2015

Um último aspecto relevante dentro os definidos como sendo os que melhor caracterizam um lojista é a diversidade de setores existentes na mercearia (Gráfico 7). Entender quais lojas possuem frios, FLV (Frutas, Legumes e Verduras), padaria, açougue e lanches é um ponto bastante relevante para entender o perfil de compra destas lojas. Por se tratar de uma característica estratégica para a atração de diferentes perfis de clientes e para o estímulo de diferentes comportamentos de compra, é fundamental que a empresa compreenda a composição das lojas de seus clientes para melhor selecionar o sortimento de produtos e a variedade de serviços que serão oferecidos no intuito de melhor atendê-los.

Gráfico 7: Distribuição dos varejistas por setores existentes na loja



Fonte: Pesquisa de Minimercados no Brasil, 2015

Uma vez definidas as características destes clientes identificados no Varejo de Proximidade, é também importante traçar um contexto a respeito das dores enfrentadas por eles em suas rotinas. Estas dores podem ser desde problemas operacionais até problemas estratégicos de posicionamento do seu negócio e serão dedicados breves parágrafos abaixo para a descrição dos principais destes problemas. Para a compilação e priorização das dores dos clientes levantadas nesta introdução foram utilizadas a experiência do autor no segmento, a convivência com outros funcionários da empresa que possuem anos de experiência no varejo alimentar e visitas de campo realizadas aos clientes da organização.

A gestão de estoques é um problema consistente entre os lojistas, de modo que fatores como a inexistência de ferramentas de controle, a ruptura de produtos e a ocorrência de perdas por vencimentos ou avarias são bastante comuns neste segmento. O controle das vendas também é muitas vezes impreciso, haja vista que a precificação muitas vezes é realizada com base apenas na experiência do dono do estabelecimento, não há um registro preciso de saídas e os produtos não são expostos da maneira mais eficiente em suas lojas. Ainda, no que se

refere aos *stakeholders* do dono da loja, este precisa em geral distribuir muito do seu tempo se deslocando até fornecedores, recebendo e atendendo vendedores em sua loja, capacitando e gerindo funcionários e atendendo seus clientes.

Tendo em vista os problemas supramencionados, muitos dos lojistas pertencentes ao perfil descrito nas categorizações do mercado de atuação e do perfil de cliente possuem dificuldades em manter a saúde financeira do seu negócio estável. E é nesse contexto que entra a empresa cliente do projeto desenvolvido neste trabalho, que será a seguir melhor caracterizada.

1.1.3 Empresa

Para fins de preservar informações eventualmente sensíveis, bem como estratégias de negócio da empresa cliente do projeto, o nome da empresa será substituído pelo nome fantasia Varejo ABC ao longo de todo o desenvolvimento deste relatório. Desse modo, onde for encontrada referência ao nome Varejo ABC, refere-se à empresa real utilizada pelo autor para este estudo de caso.

O projeto desenvolvido ao longo deste trabalho ocorre em consonância com as demandas existentes no ambiente de trabalho do autor, a Varejo ABC. Desse modo, compete a esta seção uma breve introdução sobre a empresa em questão de maneira que seja possível compreender como esta se posiciona no cenário do Varejo Tradicional analisado na seção 1.1.1 e como esta se construiu em torno do perfil de mercearias descrito na seção 1.1.2. O leitor será então apresentado aos aspectos gerais da empresa, ao grau de maturidade desta frente ao seu estado atual de rodadas de levantamento de capital, ao posicionamento da Mercê frente a benchmarks internacionais e a concorrentes locais, bem como ao papel do autor dentro da organização.

A Varejo ABC surgiu visando atender às dores do perfil de Mercearias de 1 a 4 checkouts de modo a posicioná-las de maneira mais competitiva no mercado frente aos seus concorrentes de supermercados, hipermercados e cash & carries. Os fundadores da empresa objetivaram com sua criação a inserção de tecnologia neste mercado tão tradicional em que a gestão das lojas é muitas vezes realizada de maneira subjetiva e intuitiva pelos seus donos.

Para este objetivo, a Varejo ABC visa a atender um mercado com três principais *stakeholders* tidos como clientes de suas soluções: as mercearias - já descritas em certa profundidade, os distribuidores e indústrias que fornecem os insumos para estas mercearias e eventuais provedores de serviços financeiros e tecnológicos que podem ser conectados a estes lojistas.

Para sustentar esse posicionamento de mercado, a Varejo ABC, assim como grande parte das startups nacionais, pauta-se em alianças estratégicas com investidores por meio de rodadas de investimento para sustentar sua proposta de valor. Ao longo dos anos desde sua fundação, a empresa já captou as rodadas *Seed, Seed 2 e Series A*, sendo esta a mais recente e expressiva realizada em setembro de 2021. Nesta última rodada, foi captado um investimento de $10 milhões, o equivalente a cerca de R$53 milhões.

O modelo de negócio da Varejo ABC foi desenvolvido a partir de referências internacionais de benchmarkings em empresas cujo modelo já tem sido implementado e validado ao redor do mundo nos últimos anos. A principal referência para o desenvolvimento do modelo, bem como a empresa mais expressiva do segmento no cenário global é a Ling Shou Tong na China. A empresa foi fundada pela Alibaba, uma das empresas mais relevantes na economia chinesa, e tem sido uma das maiores responsáveis pela sustentabilidade financeira das mercearias de bairro bem como pelo aumento da sua competitividade no mercado chinês através de soluções tecnológicas que atacam diretamente as maiores dores deste perfil de cliente. Além deste exemplo, existem modelos de negócio bastante similares em diversos outros países como na Colômbia (Chiper), na França (Ankorstore), na Índia (Khatabook) e no Vietnam (Telio).

No cenário nacional, não existem ainda empresas concorrentes que atuem completamente no modelo de negócios objetivado pela Varejo ABC, mas existem diversas empresas emergentes que podem ser vistas como concorrentes principalmente na parte inicial da proposta de valor da organização que é o alicerce deste modelo de negócio. Sendo assim, podem-se destacar alguns distribuidores como a Bate Forte e a Comercial Esperança e algumas outras empresas do segmento de varejo alimentar como a Frubana, a BEES e a Yandeh que são pautadas em soluções tecnológicas como os principais concorrentes da Varejo ABC neste mercado.

Neste contexto, o autor atua como parte do corpo de funcionários da empresa desde fevereiro de 2021, já tendo passado pelo setor comercial da empresa atuando no

relacionamento com parceiros da Indústria, de Distribuidores e de Atacarejos. Atualmente, Cauê Luna da Silva atua como Analista de Dados, desempenhando funções relacionadas à gestão de indicadores, à construção de modelos analíticos e à integração de plataformas.

Dado este contexto, é possível entender inicialmente a empresa cliente do projeto e seu posicionamento frente ao mercado, bem como o escopo de atuação do autor na organização. Assim, avançaremos nossa introdução para analisar em maior detalhes a proposta de valor da empresa para criarmos base para delimitarmos melhor o escopo do projeto.

1.1.4 Estratégia e Proposta de Valor

Para caracterizar a proposta de valor da Varejo ABC frente ao mercado de atuação e ao perfil de cliente expostos, será feito uso do conceito de Missão, Visão e Valores analisado por Collins e Rukstad na *Harvard Business Review* (COLLINS, 2018). Nessa análise, podemos caracterizar este tripé de fundamentos de modo que a Missão representa o porquê da empresa existir, os Valores representam as crenças da empresa e as diretrizes que determinam o comportamento dos funcionários e a Visão é o que a empresa deseja ser.

A Missão da Varejo ABC é melhorar a vida das pessoas em suas comunidades através do acesso para pequenos empreendedores e seus clientes. Esta definição reforça o comprometimento das iniciativas da empresa com o fornecimento de serviços e produtos para estes clientes com um objetivo claro de melhorar a vida no entorno de sua comunidade.

No que se refere aos valores, a empresa possui seis pilares que são o que norteiam a conduta de seus funcionários e, portanto, devem nortear também as iniciativas desenvolvidas ao longo deste projeto. Nesta breve introdução, não compete discorrer longamente acerca destes valores, mas compete destacá los a seguir:

1. “Pessoas em primeiro lugar”
2. “Foco no cliente e na comunidade”
3. “Alto Impacto”
4. “Comprometimento de Dono”
5. “Verdade e Dados”

6. “Integridade”

Por fim, em termos de visão, a Varejo ABC deseja revolucionar o mercado de proximidade através de tecnologia e soluções integradas (verticalizadas) para os lojistas e seus clientes.

Além dos aspectos analisados na *Harvard Business Review*, compete apresentar ao leitor o Sonho Grande da empresa. Este pode ser definido como: “Ser a maior rede de lojas varejistas do Brasil até o final de 2024, transformando a vida dos pequenos varejistas e de seus clientes através da tecnologia”. Neste aspecto ressalta-se a fundamentação da importância do projeto elaborado pois, para que seja possível alcançar um objetivo tão desafiador, é fundamental que sejam implementadas iniciativas tecnológicas de engenharia para garantir esta escala.

Uma vez introduzidos ao perfil de cliente e ao MVV da empresa, para entendermos corretamente o posicionamento de mercado buscado por ela é importante compreendermos também a Proposta de Valor da organização no mercado analisado. Como exposto por Alexander Osterwalder em seu livro Value Proposition Design: How to Create Products and Services Customers Wants: *“O Canvas de Proposta de Valor tem dois lados. Com o Perfil do Cliente nós clarificamos nosso entendimento. Com o Mapa de Valor você descreve como você pretende criar valor para seu cliente. O Market Fit é encontrado quando um lado encontra o outro”* (OSTERWALDER, 2018).

Dada a definição deste conceito, podemos compreender que a proposta de valor de mercado da Varejo ABC se concentra no oferecimento de soluções integradas verticalizadas para donos de mercearias e seus clientes. Este conceito pode ser melhor compreendido de modo que se interprete por soluções integradas verticalizadas todas as soluções que sejam focadas exclusivamente no perfil do cliente analisado e possam ser integradas de modo a transformar de ponta a ponta a jornada do usuário de seus serviços e também de seus clientes. Com isso, é possível compreender o portfólio de produtos e de serviços da empresa e como estes se encontram dentro da proposta de valor da organização.

1.1.5 Portfólio de Produtos e Serviços

Em um estado de desenvolvimento pleno de sua estrutura organizacional, é esperado que a Varejo ABC seja capaz de fornecer soluções integradas ao longo de toda a jornada do lojista para gerir a sua mercearia. Entretanto, no momento atual ainda há um foco inicial em conquistar o mercado em uma parte essencial do negócio de mercearias e que funciona como alicerce para as demais soluções tecnológicas futuras, que é o abastecimento de produtos.

O abastecimento de produtos pode ser compreendido como a compra de mercadorias para sua loja, que futuramente serão vendidas para seus clientes. Hoje é comum que os lojistas tenham que lidar com uma quantidade muito grande de fornecedores classificados entre indústrias, distribuidores e atacarejos para compor seu estoque de produtos.

O serviço principal da Varejo ABC - e que será foco do projeto desenvolvido neste trabalho - é a atuação como um e-commerce em que lojistas da Região Metropolitana de São Paulo podem acessar seu aplicativo e comprar uma grande variedade de produtos. O abastecimento de produtos da empresa é pautado em três pilares de qualidade: disponibilidade de sortimento, preço competitivo e tempo de entrega rápido.

Para a concretização deste serviço, existem *stakeholders* fundamentais para este trabalho que são a força de vendas - responsável por captar novos clientes e por impulsionar as vendas da plataforma para estes - o time de logística - responsável pelas entregas - o time comercial - responsável pelas negociações com fornecedores e pela seleção do sortimento de produtos ofertados - e o time financeiro - responsável pela sustentabilidade financeira das operações. Dessa forma, é possível compreender introdutoriamente como a empresa se posiciona oferecendo soluções para seus clientes para que seja possível definir adequadamente o escopo deste projeto.

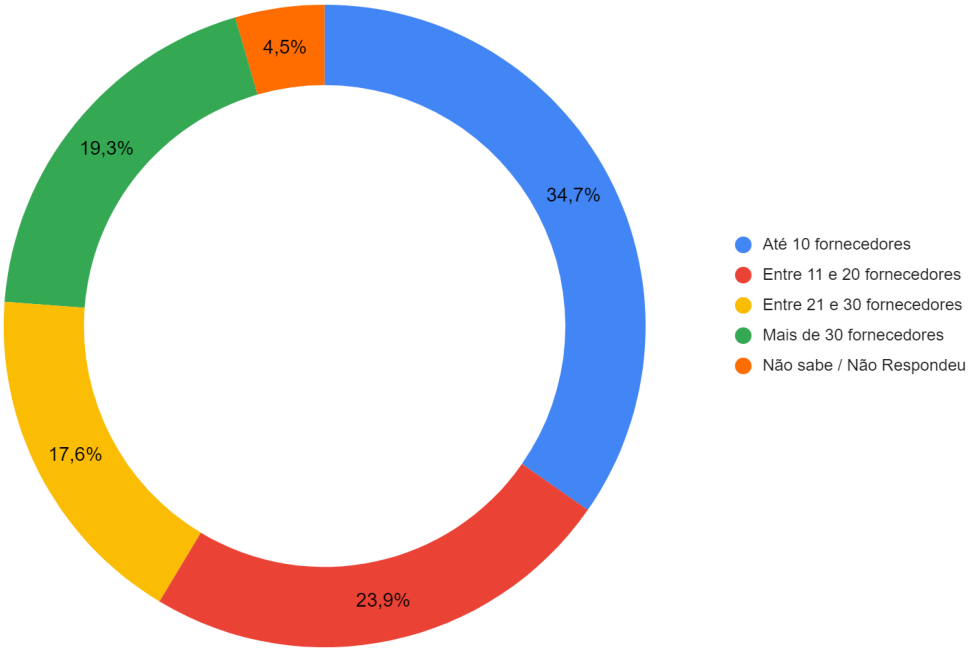
1.2 Motivações

A motivação para a execução deste projeto parte da análise do ambiente competitivo da empresa que pode ser analisado a partir dos contextos fornecidos nas seções anteriores. A

principal referência que norteia o estudo do ambiente competitivo neste trabalho são as forças competitivas observadas na análise estrutural da indústria realizada no livro *Towards a Dynamic Theory of Strategy* (PORTER, 1991). Para um bom posicionamento no mercado de abastecimento de mercearias, a Varejo ABC precisa entender as ameaças advindas do poder de barganha dos fornecedores, do poder de barganha dos consumidores, das ameaças de novos ingressantes no mercado e das ameaças de novos produtos ou serviços substitutos. Apesar de ser notória a relevância de todas estas quatro Forças de Porter para a empresa, o projeto em questão é principalmente motivado por duas delas: o poder de barganha dos consumidores e a ameaça de novos ingressantes no mercado.

Vale ressaltar que a proposta de valor da empresa não se limita ao fornecimento de mercadorias, mas que esta é uma parte fundamental do negócio pois servirá como base para o desenvolvimento de suas soluções tecnológicas verticalizadas. Sendo assim, apesar de não ser necessariamente neste ponto da sua proposta de valor que se concentra seu maior diferencial competitivo de longo prazo, a empresa precisa se posicionar competitivamente frente a seus concorrentes. Como apresentado na Pesquisa de Minimercados no Brasil (SEBRAE, 2015), os lojistas desse perfil de cliente possuem uma grande gama de fornecedores que os atendem. Sendo assim, é imprescindível que sejam identificados os lojistas com maior potencial de compra dos produtos de abastecimento fornecidos pela Varejo ABC e que sejam desenvolvidas estratégias para minar o poder de barganha dos consumidores de forma a evitar que este aspecto seja prejudicial para a sustentabilidade financeira do modelo de negócio.

Gráfico 8: Distribuição dos varejistas tradicionais por número de fornecedores



Fonte: Pesquisa de Minimercados no Brasil, 2015

Como mencionado anteriormente, apesar de existirem alguns competidores no mercado de abastecimento, não existem ainda empresas posicionadas estrategicamente no modelo de negócios similar ao da Ling Shou Tong no Brasil. Esse fato torna imprescindível que para se posicionar adequadamente nesse mercado a Varejo ABC deve estar muito próxima dos clientes com maior potencial de aderência para suas soluções integradas verticalizadas que serão complementares à proposta de valor de abastecimento. A concretização desta seleção de clientes e sua respectiva aquisição e rampagem dentro da base de clientes da empresa fará naturalmente com que sejam elevadas as barreiras de entrada deste mercado para novos competidores, o que tornará o ambiente competitivo mais promissor à cliente do projeto.

Além do posicionamento focado nas Forças de Porter priorizadas, é fundamental que a empresa se posicione adequadamente frente às dimensões da estratégia competitiva. Como abordado no livro Estratégia Competitiva (CARVALHO e LAURINDO, 2010), a amplitude das diferenças estratégicas da indústria depende da natureza da indústria. E a definição da estratégia competitiva incide em trade-offs que modelam a empresa. Sendo assim, é importante que sejamos capazes de entender quais aspectos da estratégia competitiva devem

ser foco das iniciativas da empresa para que haja um maior potencial de aquisição, retenção e aproveitamento do potencial dos clientes com elevada aderência a perfis mais promissores tanto para o consumo de produtos no modelo de abastecimento quanto para a aderência às soluções integradas verticalizadas futuras.

Outro fator determinante para a motivação à execução deste projeto pela empresa é a necessidade de captação de rodadas de investimento futuras para garantir a escalabilidade de seu modelo de negócios. Após o Series A coletado em setembro de 2021, a empresa possui recursos suficientes para a manutenção de sua operação por alguns anos. Entretanto, a implementação de novas soluções complementares à sua estratégia de abastecimento e a escala de suas operações demandam elevados recursos financeiros que deverão ser captados futuramente em uma rodada de *Series B*.

Para que esta rodada aconteça de modo a ser possível levantar uma quantidade elevada de capital sem que haja grande dissolução das ações da empresa para investidores, é fundamental que a Varejo ABC possua métricas estratégicas com uma ótima performance durante as negociações desta rodada de investimentos. Essas métricas são cuidadosamente acompanhadas pelas lideranças da empresa juntamente ao board de investidores e passam essencialmente pelo crescimento da empresa em relação à sua base de clientes e ao seu faturamento e pela sustentabilidade econômica da empresa. Sendo assim, é indispensável que sejam traçadas estratégias para a aquisição de novos clientes de modo que estes possuam elevada adesão à proposta de valor da empresa, unindo grande recorrência de compra a uma elevada margem de lucro para a empresa.

Um último ponto que motiva a execução deste projeto é a necessidade de gerir de maneira eficiente os recursos financeiros da empresa ao longo da trajetória rumo às conquistas pautadas nos demais aspectos desta seção motivacional. A área de Ciência de Decisões da qual o autor faz parte na empresa é responsável por garantir que sejam tomadas decisões de negócios que zelem pelas métricas de sustentabilidade financeira do business e que permitam a escala de maneira exponencialmente proporcional à mão de obra, usando para isso diversos fundamentos da análise de dados e da construção de modelos matemáticos.

Em resumo, este trabalho pauta-se em cinco principais aspectos motivacionais: a necessidade de redução do poder de barganha dos consumidores, a demanda por iniciativas para aumento das barreiras de entrada neste mercado para futuros concorrentes, a inevitabilidade da necessidade de realizar uma priorização de dimensões estratégicas a serem focadas neste mercado, a imprescindibilidade de novas rodadas de investimento com alto capital levantado e baixa diluição de ações e a necessidade de gerir bem os recursos

financeiros da empresa neste processo. Haja vista estes aspectos, é possível focar de maneira mais precisa em se definir o problema a ser atacado pelo autor deste trabalho.

1.3 Problema

Para a correta definição do problema central a ser atacado no desenvolvimento deste trabalho foram realizadas diversas reuniões com a área cliente do projeto, que é a área de vendas da empresa, além de reuniões com o líder da área de Ciência de Decisões e de reuniões com o *Business Owner* da empresa. Além da interface com os *stakeholders* do projeto mencionados, foram também realizadas pesquisas acerca do contexto em que o problema está definido e levantamentos bibliográficos para melhor materialização das dificuldades encontradas na empresa em um problema materializável e concretamente atacável.

O problema da empresa - dadas as motivações para o projeto apresentadas acima - está centrado na inexatidão da definição de quais são os clientes com maior adesão potencial à proposta de valor da empresa no longo prazo. Não apenas é perceptível que a empresa ainda não possui uma carteira de clientes otimizada com relação à adesão à sua proposta de valor, como também é possível compreender que ainda existe certa ambiguidade dentro na organização na definição do que de fato compreende um cliente com elevada adesão à proposta de valor.

Por conseguinte, é fundamental definir esse conceito antes da materialização do problema encontrado. Após as análises e reuniões realizadas, o autor e os *stakeholders* envolvidos chegaram à conclusão de que a melhor forma de definir a adesão do cliente à proposta de valor da empresa seria: a agregação entre a receita bruta mensal, a margem de bruta percentual e a variedade de skus consumidos mensalmente pelo cliente. Por consequência, quanto maior a performance do cliente nestes indicadores, maior seria a adesão do cliente à proposta de valor deste no modelo de negócios.

Em um cenário utópico, existiriam dados suficientes para saber exatamente tudo que um cliente compra para abastecer a sua loja, bem como a frequência e o ticket médio dos pedidos que este realiza. Entretanto, em um cenário factual sabe-se que a análise populacional para um universo de mais de 400 mil mercearias no território nacional seria impossível tanto pelo processo de coleta destes dados quanto pela falta de acesso a eles, dado que o cliente compra

de diversos fornecedores além da Varejo ABC, o que gera uma grande gama de dados que não podem ser facilmente acessados.

Tendo em mente essas limitações, o problema corrente se encontra no fato de que atualmente não é sabido sequer o verdadeiro potencial de compra do cliente dentro do e-commerce da Varejo ABC. Esta lacuna de conhecimento estratégico dificulta muito a capacidade de se atuar de maneira estratégica na captação, retenção e exploração de clientes consumidores potenciais na base de clientes da empresa.

Outrossim, além de não sabermos atualmente qual o potencial de compra de um cliente, não sabemos quais são suas características intrínsecas que aumentam ou diminuem a adesão desse cliente à proposta de valor da empresa. Sendo assim, torna-se muito mais complexa a tarefa de filtrar quais clientes devem ser prospectados pela área de vendas, quais clientes devem ser removidos da base atual e quais clientes deveriam possuir ações direcionadas a mudar seu perfil de compra de modo a otimizar a eficiência econômica do modelo de negócios para os patamares necessários a uma gestão eficiente de recursos e a uma captação de boas rodadas futuras de investimento.

Ademais, é também notório que a empresa - além de desconhecer a aderência efetiva de cada cliente e as características que a impulsionam - ainda poderia ter formas muito mais elaboradas e personalizadas para aproveitar melhor o potencial de compra desses clientes. Estas formas poderiam ser focadas tanto na captação de novos clientes com elevado potencial, quanto no filtro da base atual de clientes para manutenção apenas de clientes com potencial acima do desejável, passando também por iniciativas voltadas a impulsionar as vendas aos clientes que estão performando abaixo do potencial esperado e a compreender o estudo de caso de clientes com performance acima do potencial esperado.

1.4 Objetivos

Em face dos problemas mencionados, o problema-alvo deste projeto pode ser definido de maneira extremamente objetiva. Sendo assim, este objetivo possui em sua definição quatro vertentes auxiliares:

1. Definir e caracterizar os perfis de compra dos clientes existentes hoje na organização;
2. Determinar a composição ideal da base de clientes entre os perfis definidos;
3. Compreender o processo de transição de clientes entre perfis de compra;

4. Identificar os atributos com maior influência no perfil de compra de um cliente.

Em suma, esses quatro objetivos auxiliares serão partes essenciais do desenvolvimento do objetivo primordial do projeto que, fazendo uso dos princípios de definição de um objetivo *S.M.A.R.T.* (DORAN, 1981), pode ser descrito como:

"Consolidar, até o final do ano de 2022, planos de ação implementáveis que sejam capazes de alterar a distribuição dos perfis de cliente da base atual da Varejo ABC para uma composição que otimize a adesão desta base à proposta de valor da empresa"

1.5 Estrutura do Trabalho

O presente trabalho está estruturado em cinco sessões, sendo este o fim da primeira delas, a introdução. Na etapa seguinte do trabalho, será feita uma revisão da literatura. Durante este processo espera-se poder retomar alguns dos conceitos mais relevantes no âmbito da Engenharia de Produção e de outras áreas do conhecimento interdisciplinares que serão utilizados neste trabalho a fim de consolidar o embasamento teórico para a aplicação prática das análises desenvolvidas para a empresa cliente.

Tendo já se familiarizado com as etapas iniciais que dizem respeito à introdução e à revisão bibliográfica, o leitor será apresentado à metodologia do projeto. Nesta seção a metodologia utilizada pelo autor será descrita em maiores detalhes a fim de elucidar a utilização dos conceitos apresentados na bibliografia deste material no desenvolvimento. Ainda, serão apresentadas as ferramentas utilizadas na coleta dos dados a serem utilizados, serão apresentados os dados resultantes desta coleta e será feita uma breve análise sobre estes para nortear a sua utilização.

Em seguida, nos resultados do projeto serão apresentados os modelos construídos e as iniciativas traçadas durante os meses da implementação do projeto na empresa. Ao longo da seção, os resultados práticos serão compartilhados com o leitor bem como as observações e constatações realizadas nas reuniões de avaliação do projeto com os *stakeholders* envolvidos.

Por fim, serão compiladas as conclusões acerca do projeto desenvolvido em uma seção final do trabalho desenvolvido. Neste contexto, serão também sumarizadas as principais contribuições do conteúdo elaborado para a empresa cliente e para a ciência através da publicação deste material.

2 REVISÃO DA LITERATURA

A revisão da literatura que compõe este documento foi construída de maneira gradual ao longo de todo o desenvolvimento deste, de modo que foi revisitada durante muitos momentos desde as etapas mais introdutórias até as etapas finais do projeto. Assim, faz-se possível que a literatura utilizada esteja sempre atualizada e que o trabalho desenvolvido esteja sempre embasado em uma literatura sólida bem revisada pelo autor e que possa ser revisitada a qualquer momento durante a leitura do trabalho.

Esta seção divide-se primordialmente em cinco subseções sendo elas: métodos de abordagem para o problema proposto, ferramentas utilizadas, conceitos de contabilidade, métodos de clusterização e métodos de correlação. Cada uma destas subseções será composta por um ou mais conceitos referenciados na bibliografia deste documento que serão revisados nesta seção e posteriormente implementados no projeto nas seções seguintes.

2.1 Métodos de abordagem para o problema proposto

O projeto pressupõe a adequação de seus desenvolvimento à estrutura da empresa e à cultura da empresa na qual este foi desenvolvido. Sendo assim, serão revisitados três conceitos bastante comuns no contexto da Engenharia de Produção e do desenvolvimento de projetos. Primeiro, para garantir adequação da metrificação dos resultados do projeto à estrutura de metas e indicadores da empresa, será revisado o conceito de Objetivos e Resultados-Chave. Em seguida, visando a adequação à cultura de desenvolvimento de projetos da organização, será abordado o conceito de *Lean Startup*. Será também abordado o Princípio da Pirâmide, que será utilizado amplamente durante as etapas de planejamento e de apresentação de resultados do projeto.

2.1.1 Objetivos e Resultados-Chave

Desenvolvidos por Andy Groove no ano de 1971 enquanto trabalhava na Intel e descritos por John Doerr no livro Avalie o que importa (DOERR, 2019), os Objetivos e

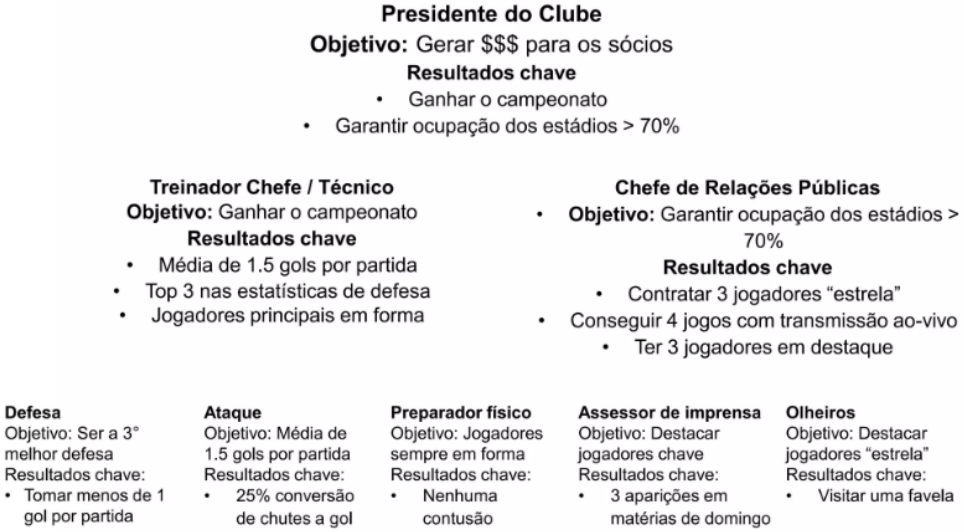
Resultados-Chave (*OKRs*) é um protocolo colaborativo para definição de metas para equipes, empresa e indivíduos. A metodologia consiste basicamente em objetivos, que são O QUE deve ser alcançado e em Resultados-Chave, que estabelecem e monitoram COMO o objetivo deve ser alcançado. Os objetivos, por definição, devem ser significativos, concretos, orientados por ações e inspiradores. Já os Resultados-Chave são específicos, limitados no tempo, agressivos, porém realistas.

Esta metodologia gera quatro principais benefícios, descritos por John Doerr como superpoderes, que são as principais razões para sua implementação em uma organização. O primeiro destes superpoderes é o foco e comprometimento com as prioridades, haja vista que os OKRs forçam os líderes da empresa em tomar decisões difíceis com relação à priorização dos indicadores que devem ser focados pela empresa. Em seguida, esta metodologia também permite um aumento do alinhamento e da conexão do trabalho com as equipes, pois possui uma estrutura de relação bastante intensa entre os *OKRs* de responsabilidade de cada indivíduo na empresa de forma que todos tendem a mirar em objetivos comuns convergindo seus esforços. Em terceiro lugar, esse protocolo de definição de metas assegura melhor rastreamento das responsabilidades entre as áreas e entre os colaboradores, de modo que quando algum objetivo não está sendo alcançado, fica evidente toda a cascata de responsabilidades. E, por fim, os *OKRs* também permitem às empresas buscarem resultados surpreendentes, tendo em vista que por definição são extremamente agressivos e desafiadores, porém alcançáveis.

A correta implementação dos Objetivos e Resultados-Chave dentro de uma empresa depende essencialmente de três fatores. A implementação de *CRFs* (Ciclos de Conversa, Reconhecimento e *Feedback*) através dos quais os colaboradores se mantém constante alinhados através da prática desse meio de gerenciamento de desempenho. Ademais, é fundamental a implementação de uma cultura de melhoria contínua na empresa para a evolução tanto do planejamento e definição dos *OKRs* quanto para seu correto acompanhamento e atingimento ao longo dos ciclos. A cultura organizacional - que é definida por Edgar Schein como um conjunto de pressupostos básicos que um grupo inventou, descobriu ou desenvolveu para lidar com problemas de adaptação externa e integração interna e que funcionaram bem o suficiente para serem considerados válidos e ensinados a novos membros como forma certa de enfrentar esses problemas (SCHEIN, 2009) - é o último dos três aspectos essenciais para a garantia da boa funcionalidade da metodologia em um organização.

Na sequência, é possível ver um exemplo de OKRs (Ilustração 1) descrito no livro de John Doerr que ilustra a aplicação da metodologia no contexto de um time de futebol americano. O exemplo traz claramente a definição de Objetivos e de Resultados-Chave que seguem as características descritas e que possuem uma boa estrutura de cascateamento entre os envolvidos.

Ilustração 1: Exemplo de Aplicação de OKRs em um time de futebol americano



Fonte: Avalie o que importa (2019)

2.1.2 Lean Startup Enxuta

O conceito de Lean Startup Enxuta foi desenvolvido por Eric Ries em 2011 e é uma das mais importantes literaturas contemporâneas sobre o universo das emergentes Startups. No livro que dá nome ao conceito, o autor descreve uma série de experiências práticas que teve ou analisou no intuito de assegurar o embasamento para a definição dos conceitos da obra. O método *Lean Startup* possui cinco princípios fundamentais que são de necessário entendimento para o desenvolvimento de projetos em uma organização que faça uso deste.

Inicialmente, o método define que empreendedores estão em todos os locais, de modo que qualquer organização que precise inovar necessita de empreendedores e define uma Startup como sendo: uma instituição humana com o objetivo de criar novos produtos e

serviços sobre condições de extrema incerteza. Na sequência, afirma que, com base no primeiro princípio que empreender é, para além do conceito de administração tradicional, administrar sob condições de extrema incerteza (RIES, 2011).

A aprendizagem validada é um dos princípios de maior foco dentro da obra de Ries e discorre sobre a importância de se validar cientificamente - e por meio de experimentos que permitam aos empreendedores testar suas hipóteses - os aprendizados de uma organização. O ciclo construir-medir-aprender toma posse da posição de quarto princípio do método Lean Startup haja vista que o método pauta-se na capacidade dos empreendedores de usarem o ciclo de feedback para medir as reações dos clientes e para aprender se é necessário pivotar ou perseverar.

Por fim, Eric Ries discorre sobre a contabilidade para inovação, reforçando a importância de se adotarem novos métodos de indicadores para empresas que operam sob condições de extrema incerteza e com potenciais exponenciais de crescimento que vão além da contabilidade tradicional. Com os cinco princípios supracitados em mente, tem-se uma base sólida para a adaptação do projeto em curso à realidade de uma empresa que tem a cultura Lean intrínseca a ela.

2.1.3 O Princípio da Pirâmide

Um método eficiente para o planejamento do trabalho e também para a estruturação das apresentações a serem realizadas aos *stakeholders* envolvidos no projeto é o Princípio da Pirâmide de Minto, que foi desenvolvido pela autora que deu nome ao método, Barbara Minto, em seu livro *The Minto Pyramid Principle: Logic in Writing, Thinking and Problem Solving* (MINTO, 2010).

Neste método as ideias apresentadas são expostas em uma estrutura similar à de uma pirâmide, em oposição ao estilo padrão de narrativa que ocorre no sentido inverso de uma pirâmide. Em uma narrativa, o narrador inicia fornecendo contexto e gerando um clímax, passa por detalhes e centraliza a informação de forma sequencial, sequência essa que funciona muito bem para entretenimento de uma narrativa pois acaba geralmente com o clímax. Já no estilo da pirâmide, a comunicação inicia-se com a resposta apresentada, seguida pela

persuasão lógica de forma clara e explícita e é extremamente eficiente para a comunicação de decisões no contexto corporativo. O princípio da pirâmide precisa de três etapas fundamentais para seu desenvolvimento.

Inicialmente, faz-se necessário definir a pergunta-chave que embasa o objetivo da comunicação e para tanto é utilizado o método SCQ (Situação - Complicação - Questão). A resposta para esta pergunta deve ser clara, acionável e não ambígua.

Em seguida é necessário destrinchar a pergunta-chave em sub-problemas para serem atacados sequencialmente. Cada um desses problemas deve ser destrinchado em sub-problemas menores até que se atinja a causa raíz de cada um destes em uma metodologia bastante similar à implementação dos 5 porquês (OHNO e BODEK, 1988) para cada um destes. Cada um dos problemas que compõem a pirâmide descrita devem ser claramente diferenciáveis dos demais de modo a não haver lacunas nem sobreposições na definição do problema inicial exposto pela pergunta-chave, ou seja, a definição da pirâmide deve ocorrer de modo MECE - (Mutuamente Excludente e Coletivamente Exaustivo).

Por fim deve-se fazer uma série de validações na estrutura da pirâmide: verificar se a pergunta-chave está respondida, verificar se esta está disposta em uma ordem vertical clara, garantir que cada nível da pirâmide resuma perfeitamente as ideias presentes no nível anterior a ela e garantir que em cada grupo de ideias todas as presentes sejam do mesmo tipo. Caso haja falhas nesta etapa, deve-se ajustar a pirâmide construída até o estágio de perfeição.

Este framework será utilizado ao longo do desenvolvimento deste projeto tanto para a organização das ideias que visem responder o objetivo inicial dele quanto para organizar as apresentações de resultados tanto para fins de documentação neste relatório quanto para fins de apresentação aos *stakeholders* envolvidos.

2.2 Métodos Contábeis

Como definido inicialmente, a margem bruta da empresa será um dos fatores de maior relevância para a análise dos perfis de clientes ao longo do projeto. Para isso, será necessário adotar uma metodologia de cálculo base para a margem bruta retomando conceitos contábeis.

Para este fim será utilizado a DRE (Demonstrativo de Resultados) por padrão da organização, que utiliza este método para as apurações contábeis dos resultados obtidos.

2.2.1 Demonstrativo de Resultados

A metodologia do de Demonstrativo de Resultados consiste na síntese de todos os registros de receitas, custos e despesas incidentes em um determinado período na organização (INVESTOPEDIA, 2022). No caso da Varejo ABC, a DRE é elaborado mensalmente para o fechamento de contas da empresa com a contabilidade e é uma ferramenta fundamental para o registro da capacidade de empresas de gerarem ou não lucro, o que é fundamental para a sustentabilidade e para, no caso de startups, levantar boas rodadas de investimento para o financiamento das operações enquanto estas ainda não são lucrativas. Geralmente o Demonstrativo de Resultados é, para muitas empresas, umas das três principais formas de controle contábil das organizações juntamente com o Balanço Patrimonial e com o Fluxo de Caixa.

Na Demonstração de Resultados, podemos interpretar cada uma das linhas do demonstrativo a fim de entendermos adequadamente a composição do indicador de margem bruta a ser analisado no projeto. Para a obtenção do Lucro Bruto, é necessário subtrair-se da Receita Operacional Líquida o Custo das Mercadorias Vendidas, que compreende o valor pago pelos fornecedores por estas e no qual podem também ser considerados créditos de impostos que são benefícios fiscais comuns em algumas categorias de produtos no mercado de Varejo Alimentar. Para além do Lucro Bruto, nas linhas seguintes pode-se deduzir as Despesas Operacionais da empresa, que podem ser em três categorias: Custo Logísticos, Custos de Vendas e Marketing e Custos Financeiros.

Os Custos Logísticos são todos os custos de armazenagem, movimentação e transporte dos produtos em toda sua cadeia, indo desde a coleta dos produtos em fornecedores até a entrega dos produtos aos clientes e passando também por diversas linhas de custos incidentes no estoque da organização. Já os Custos de Vendas e Marketing são todos os custos que uma empresa despende para fazer com que seus clientes tenham acesso a mais informações sobre seus produtos para impulsionar vendas. E, por fim, os custos financeiros representam tanto os custos de inadimplência quanto os custos operacionais dos processos financeiros que são necessários para a manutenção da operação.

Tendo sido removidos todos os custos operacionais do Lucro Bruto, removem-se também os impostos sobre lucros no caso de balanços financeiros em que os lucros sejam

positivos e obtêm-se o Lucro Líquido da empresa no período (COSTA, 2009). Para este projeto, o foco estará na otimização do Lucro Bruto, sem que haja ainda foco no nível de Lucro Líquido da DRE, entretanto os custos operacionais por diversas vezes precisarão ser consultados e analisados para garantir o alinhamento das iniciativas desenvolvidas com a obtenção de clientes mais rentáveis, isto é, que possuam melhor percentual de Lucro Líquido para a empresa será de extrema importância para o sucesso do modelo de negócios desenvolvido.

2.3 Ferramentas utilizadas

Este projeto foi desenvolvido dentro do escopo da área de Ciência de Decisões da Varejo ABC e, por esta razão, fará uso das ferramentas padrões utilizadas nesta área da organização. O uso de ferramentas padronizadas será essencial para garantir a integração das entregas deste projeto com os sistemas de Tecnologia da Informação previamente existentes, o que tornará muito mais fluido o uso destas entregas pelos seus usuários. O projeto fará uso de seis ferramentas principais que são, nominalmente: *Databricks, PyCharm, Pyspark, PEP-8, PostgresSQL, Metabase*. Cada uma dessas ferramentas possui aplicações e funcionalidades distintas dentro do contexto da Tecnologia da Informação e da Ciência de Dados e será utilizada com um propósito específico dentro do projeto. Compete também elucidar que estas ferramentas terão atuação complementar ao longo do desenvolvimento das entregas, de modo que frequentemente os outputs dos materiais desenvolvidos com uso de uma ferramenta serão utilizados como inputs para o que será desenvolvido com as demais ferramentas.

2.3.1 Databricks

O *Databricks* é um *Data Lakehouse*, estrutura que unifica pontos fundamentais tanto de um data warehouse quanto de um data lake em uma única plataforma. Esta plataforma é capaz de suportar diversos modelos de dados, códigos, clusters, análises e modelos que serão fundamentais para a construção do projeto.

Para a correta utilização desta plataforma, será constantemente utilizada a documentação oficial do databricks (DATABRICKS, 2022) na qual são registradas de maneira constante todas as informações necessárias acerca de cada uma das atualizações da

plataforma. A interface do databricks armazena atualmente todos os bancos de dados da Varejo ABC que serão utilizados nas análises a serem realizadas, além de outros códigos desenvolvidos por outros funcionários da empresa que podem ser fonte de tabelas e de análises auxiliares e que poderão ser referenciadas ao longo deste projeto.

2.3.2 PyCharm

Como o *IDE* (Ambiente de Desenvolvimento Integrado) a ser utilizado neste trabalho, foi selecionado o PyCharm, ferramenta desenvolvida pela Jetbrains para desenvolvedores, que é uma das mais confiáveis e amplamente utilizadas no mercado de tecnologia na atualidade. Para o melhor aproveitamento das funcionalidades desta ferramenta, será extensivamente utilizada a plataforma de conteúdos da *Jet Brains* em que se encontram as documentações e guias da ferramenta (PYCHARM, 2022).

Nas seções seguintes, bem como nos apêndices, serão incluídas diversas capturas de tela que representarão o código em desenvolvimento nesta ferramenta. Ao longo de todo o projeto, haverá a preocupação com a utilização de nomenclaturas claras e de comentários pertinentes para possibilitar ao leitor a compreensão dos registros realizados.

2.3.3 Pyspark

Pyspark é uma interface do Apache Spark no Python através de uma biblioteca. Esta permite tanto que sejam escritas aplicações em Spark utilizando APIs do *Python* quanto suporta diversas ferramentas fundamentais para o desenvolvimento deste projeto como o *Spark SQL, DataFrames* e Bibliotecas de *Machine Learning* (SPARK, 2022).

Spark SQL é uma estrutura de processamento de dados que lida muito bem com a abstração de programação denominada de DataFrames. Esta estrutura por sua vez é fundamental para que seja possível utilizar com propriedade as tabelas que serão apresentadas na seção referente à coleta de dados e que serão os alicerces para a construção dos modelos apresentados neste trabalho.

Por fim, esta interface é essencial para a padronização das entregas de acordo com a estrutura padrão de códigos da empresa cliente, em que os desenvolvimentos da área de Ciência de Decisões são elaborados de maneira complementar utilizando *SQL* e *Pyspark* no *Python*.

2.3.4 PEP 8

Da mesma forma que haverá a preocupação do correto uso das ferramentas citadas acima para a construção do projeto, faz-se necessário um guia de estilo para o código a ser desenvolvido. Neste intuito, foi estudado o PEP 8 - Guia de Estilo para Códigos em *Python* (ROSSUM, 2001) que é o guia mais popularmente utilizado por desenvolvedores da linguagem para a padronização de seus códigos.

Dentro da empresa, há uma constante preocupação na correta documentação dos projetos desenvolvidos, bem como da perenidade dos resultados alcançados. Para isso, houve um estudo deste guia ainda nos momentos iniciais de desenvolvimento do projeto para garantir que os resultados estejam consistentes com os padrões compreendidos por outros desenvolvedores, sejam eles simultâneos ou não à permanência do autor nesta área da empresa.

2.3.5 PostgreSQL

O *PostgreSQL* é atualmente o banco de dados relacional open source mais avançado no mundo. Neste banco de dados será possível ao autor ter acesso às diferentes tabelas que compõem o banco de dados relacional da organização de modo a permitir a conexão de informações de pedidos, de clientes, de faturas, entre muitas outras que são essenciais para o desenvolvimento da empresa. Este banco de dados utiliza a linguagem de programação *SQL* para sua consulta, o que simplifica e otimiza o consumo de poder computacional em consultas de informações no banco de dados.

O banco de dados possui grande notabilidade no meio de desenvolvedores por contar com uma boa arquitetura, integridade de dados, confiabilidade e features robustas, além de ser uma plataforma de código aberto colaborativa. Assim como no caso de muitas das outras

ferramentas citadas nesta seção, o *PostgreSQL* foi estudado utilizando-se a documentação oficial (POSTGRESQL, 2022) da plataforma que contém todas as atualizações sobre features contidas nela.

2.3.6 Metabase

A última ferramenta fundamental para o projeto que precisou ter sua literatura revisada nas etapas anteriores ao início da execução do projeto é o *Metabase*. Estando já devidamente contemplados o data lakehouse do projeto, as bibliotecas *Python* a serem utilizadas, o IDE onde o projeto será desenvolvido e o guia de estilo que ordenará o desenvolvimento dos códigos em *Python*, é necessário o uso também de uma plataforma de Business Intelligence, que será o Metabase.

Para a capacitação do autor para o uso do Metabase no projeto, foi-se estudado o README da plataforma (METABASE, 2022), que é elaborado considerando instruções e documentações de todas as funcionalidades para as quais a ferramenta pode ser utilizada. Esta ferramenta possuirá três principais aplicações ao longo do projeto.

A primeira delas é fornecer acesso aos dados gerados nas análises para os *stakeholders* do projeto não pertencentes à área de Ciência de Decisões. Como muitas das ferramentas supracitadas são complexas e demandam grande estudo de suas documentações para serem utilizadas, além de não serem de acesso a todos os usuários da organização, o Metabase será o ponto final de exibição dos dados gerados, sendo assim acessível por todos na organização, inclusive para aqueles que são leigos nas linguagens de programação utilizadas.

Uma segunda aplicação da plataforma é a capacidade de construção de querys em *SQL* utilizando o *PostgresSQL* que, embora também possíveis de serem construídas no Databricks, muitas vezes possuem seu uso facilitado nesta ferramenta. E por fim, também será possível a realização de diversas construções gráficas no Metabase que poderão ser integradas aos Dashboards da empresa para acompanhamento dos resultados de performance decorrentes do projeto de maneira contínua.

2.4 Conceitos de Machine Learning

Para uso dos métodos de Machine Learning necessários a este trabalho, foi necessária a revisão de conceitos gerais essenciais para o embasamento dos modelos. Notoriamente não será discorrido sobre todos os conceitos da área do conhecimento de Machine Learning que serão utilizados nos modelos construídos na sequência, haja vista a amplitude de conceitos que embasa a construção destes, mas será discorrido sobre todos os principais conceitos que são fundamentais para que o leitor tenha o correto entendimento desta obra. Serão abordados os modelos de features que serão utilizados nos datasets do projeto, enunciado tanto features Categóricas quanto Features Ordinais que serão utilizadas para testar e para comparar os modelos criados.

2.4.1 Features Categóricas

Este modelo de features também pode ser comumente referenciado como features nominais. Para este caso, existem duas ou mais categorias em que a variável em questão pode ser classificada, de modo que não há uma ordem intrínseca entre as categorias. Este último ponto é bastante relevante pois, caso as categorias possuam uma ordenação clara entre elas, a feature na verdade é classificada como ordinal (UCLA, 2021). Um exemplo para as features categóricas é o time para qual um indivíduo torce. Neste caso, existem categorias específicas, porém não há uma ordenação adequada para estas categorias.

2.4.2 Features Ordinais

As Features Ordinais possuem como característica distintiva das features categóricas o fato de que estas podem ser claramente ordenadas. A diferença entre as distâncias dos grupos de features pode ser extremamente variada, mas desde que estas possam ser devidamente ordenadas, ainda fez-se válida a classificação ordinal destas (UCLA, 2021). Um exemplo de features ordinais é o indicador de NPS de uma empresa, em que existem diferentes classificações que um serviço pode receber, variando de 0 a 10, e estas podem ser perfeitamente ordenadas de modo que a menor seja 0 e a maior seja 10.

2.5 Métodos de Clusterização

Métodos de classificação podem ser utilizados para organizar os dados em grupos. Quanto temos grupos com uma definição clara, ou seja, grupos em que há um consenso sobre o que define os participantes deste grupo como, por exemplo, o artista preferido de uma pessoa ou o estado em que ela mora, podemos utilizar métodos convencionais de classificação como Regressão Logística, Árvores de Classificação ou Floresta Randômica que auxiliam a colocarmos cada conjunto de dados em um grupo distinto de acordo com esta característica que define o grupo.

Entretanto, existem casos em que não é possível definir esses grupos com clareza tendo como exemplo a classificação de perfis de clientes. Nessa situação é necessária a implementação de um método de clusterização, que agrupa os dados em grupos de entidades similares uma vez que não existem legendas que classifiquem os dados nestes grupos (SUNDARAM, 2020). Estes métodos são de extrema importância quando tratamos de grandes volumes de dados e queremos tomar ações referentes a estes dados porém não queremos personalizar a ação para cada data point que possuímos em nossa população e, ao invés disso, agrupamos os clientes em clusters e direcionamos ações personalizadas a cada cluster ao invés de a cada data point.

No caso deste projeto, faremos uso de métodos de clusterização para agrupar clientes quanto à adesão destes à proposta de valor da Varejo ABC, que é uma classificação imprecisa, ou seja, não existem grupos previamente definidos para esta classificação. Em geral, os métodos de clusterização podem ser definidos em quatro categorias: modelos de conectividade, modelos de centróides, modelos de distribuição e modelos de densidade (Portal Data Science, 2018).

Clusterização por conectividade se adota a premissa de que quanto mais perto se encontram pontos no espaço, maior é a semelhança entre estes pontos. Este tipo de modelo possui fácil interpretação porém não pode ser escalado para uma grande quantidade de dados. O exemplo mais notório de método de clusterização por conectividade é o método de clusterização hierárquica.

Já a clusterização por centróides consiste em um grupo de métodos que são iterativos e que tem noção de similaridade derivada da proximidade de cada data point ao centróide de cada cluster, de modo que os pontos são alocados nos clusters de maior proximidade com ele. Para a definição dos centróides, o número de clusters precisa ser conhecido previamente e, para isso, é importante que os dados trabalhados sejam bem conhecidos por quem aplica o

método. O *K-Means* é o método mais utilizado no universo da ciência de dados dentre todos os métodos deste grupo.

No contexto da clusterização por distribuição, como o próprio nome já indica, pressupõem que os dados seguem distribuição estatística, como a Distribuição Normal ou a Distribuição Gaussiana. O principal ponto negativo que este modelo tem é o fato de ter alto viés, ou seja, tem grande aderência aos dados usados no treinamento do modelo e pouca aderência aos dados usados no teste do modelo. O modelo de maximização de expectativas pode ser citado com um exemplo desta categoria.

Por fim, os modelos de clusterização por densidade são modelos que lidam muito bem com dados aninhados entre si em que a densidade dos pontos no espaço possui grande relevância, de modo que os clusters são formados em áreas de elevada densidade e outliers são classificados nas áreas de baixa densidade. Temos o *DBSCAN* como um bom método para este propósito.

Após o estudo destes quatro grupos de algoritmos de clusterização foi selecionado o grupo dos algoritmos de clusterização por centróides. Isso de dá pois os dados de clientes da empresa não necessariamente possuem relação atribuída à proximidade destes no espaço - caso no qual seria indicado o uso da clusterização por conectividade - os dados apresentados não necessariamente seguem uma distribuição estatística - que seria condição necessária para o uso dos algoritmos de clusterização por distribuição - e os clientes não estão aninhados e a densidade destes no espaço não é de grande relevância para sua classificação - que seriam condições necessárias para o uso dos métodos de clusterização por densidade.

Portanto, neste projeto será usada a clusterização por centróides de modo que a classificação de um cliente em um cluster será predominantemente definida pela sua proximidade dos dados referentes ao cliente ao centróide do cluster. Neste grupo existem diversos outros métodos que foram estudados pelo autor, como *K-NN* (K Near Neighbours) porém o mais popular na literatura e que foi o selecionado para implantação neste projeto foi o método *K-Means*.

2.5.1 K-Means

O algoritmo *K-Means* consiste em uma abordagem incremental de clusterização que adiciona um centro ao cluster por vez em um mecanismo de busca determinístico composto

por N repetições iterativas (LIKAS, VLASSIS, e VERBEEK, 2003). Esse algoritmo é direcionado para a resolução do problema do agrupamento de elementos em clusters e pode ser utilizado para diversas aplicações. No contexto deste projeto, o algoritmo será aplicado para agrupar os clientes da Varejo ABC em clusters de performance similares no que se refere ao seu perfil de compra, que será mensurada como uma combinação entre receita bruta, margem bruta percentual e variedade de skus comprados.

Dado um data set $X = \{x_1, \dots, x_n\}$, $x_n \in R^d$, o método K-Means particiona o data set em M clusters $C_1, \dots, C_M$ de forma que tenhamos uma clusterização otimizada por meio de iterações sucessivas. Para a clusterização por centróides descrita no tópico anterior é utilizada a distância entre o ponto x_i e o centróide m_k (definido como o centro do cluster) pertencente ao cluster C_k que contém x_i . Essa distância pode ser mensurada de diferentes formas, porém a mais usual é a Distância Euclidiana entre dois pontos, que pode ser definida pela Equação Euclidiana dos Mínimos Quadrados aplicada ao algoritmo K-Means (LIKAS, 2003:

$$E(m_1, \dots, m_M) = \sum_{i=1}^{N} \sum_{k=1}^{M} I(x_i \in C_k) \times \left\| x_i - m_k \right\|^2 ,$$

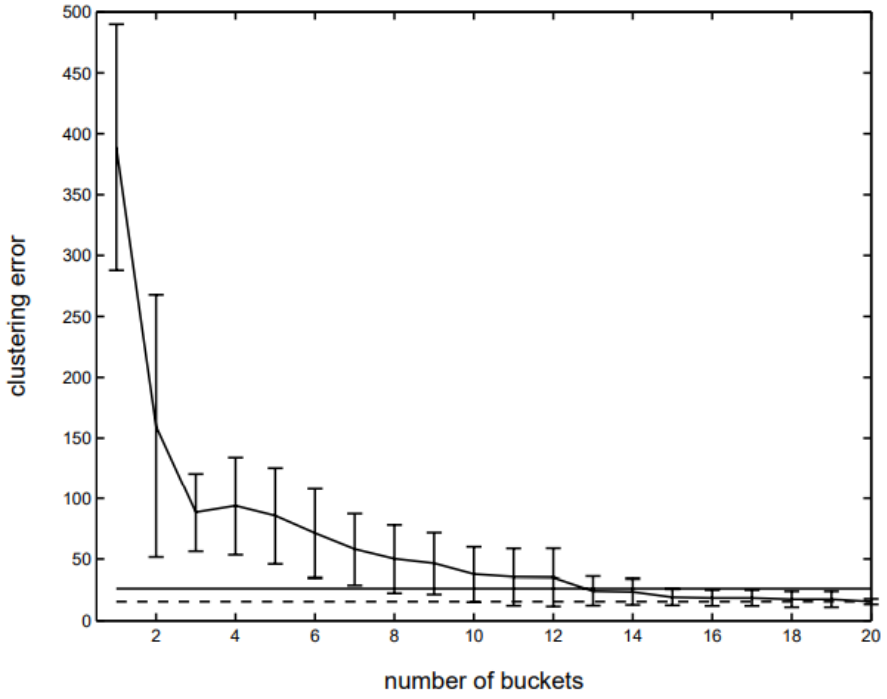
tal que $I(X) = 1$ se X for verdadeiro e X = 0 caso contrário

O algoritmo K-means atua justamente na minimização da soma dos mínimos quadrados da comparação de cada elemento com a performance da média descrita para seu cluster. Um fator importante a ser determinado para a aplicação desse método é a quantidade K de clusters na qual os dados serão particionados. Em um extremo, se adotado K=1 todos os dados estariam em um mesmo cluster e a distância euclidiana seria extremamente elevada e no outro extremo K=N cada data point teria seu próprio cluster e a distância euclidiana seria nula. Portanto, é importante definir até quando haveriam ganhos expressivos em se aumentar o valor de K, que resultaria na diminuição do erro resultante do método aplicado compensado por um aumento de esforço posterior ao se necessitar adotar ações a mais um grupo extra.

Caso se conheça perfeitamente os dados, é possível definir inicialmente o valor de K, porém também é possível utilizar o Método do Cotovelo para a definição deste valor. Neste método é aplicado o método K-Means para diferentes valores de K, mensura-se na sequência o erro da clusterização por meio do Método dos Mínimos Quadrados e plota os valores em um

gráfico que tem como eixo x o valor de K e como eixo y o erro resultante do método para o respectivo valor de K (Gráfico 9).

Gráfico 9: Método do Cotovelo



Fonte: The Global K-Means Clustering Algorithm, 2002

Uma vez definido o valor de K, o algoritmo executa os seguintes passos (STARMER, 2020):

1. Define de forma arbitrária K data points pertencentes ao dataset X
2. Mede a distância entre cada ponto não clusterizado e o centróide de cada cluster (na primeira repetição o centróide será igual à posição do ponto inicial)
3. Anexa cada ponto ao cluster cuja distância entre ele e o centróide é a menor de todas
4. Calcula a média dos pontos de cada cluster
5. Atualiza o centróide do cluster para a média calculada
6. Mede a distância de cada ponto e o novo centróide de cada cluster
7. Atualiza os pontos de modo que pertençam aos clusters cuja distância entre o ponto e o novo centróide e o ponto seja a menor de todas
8. Repete os passos 2 a 8 até que não hajam mais mudanças de cluster por nenhum ponto
9. Calcula a variância de cada cluster
10. Calcula a variância total entre os clusters

Apesar de muito significativo, este modelo representa limitações de performance pois possui elevada dependência das condições de início do método. Para contornar o problema descrito, existem duas principais abordagens que poderão ser utilizadas. A primeira delas parte de métodos de otimização global estocástica e a segunda delas parte de múltiplos resets do algoritmo. Devido ao número de data points em nosso data set não ser tão elevado, o segundo e mais simples método será utilizado de modo que será utilizado o Método de Validação Cruzada para identificar qual das versões do algoritmo será utilizada como a definitiva.

Com a aplicação deste algoritmo, a distância euclidiana entre os pontos e os respectivos clusters é minimizada, resultando na melhor clusterização possível para o dataset fornecido de acordo com esta metodologia.

2.6 Métodos de Correlação de Variáveis

A correlação é uma medida adimensional que é comumente utilizada na literatura para a análise do relacionamento linear entre pares de variáveis de diferentes unidades (JUNQUEIRA, 2018). A correlação entre duas variáveis X e Y pode ser expressa pela seguinte fórmula:

$$correlação(X,Y) = \frac{covariância(X,Y)}{desvio\ padrão(X) \times desvio\ padrão\ (Y)}$$

O termo covariância(X, Y) na fórmula representa a medida do relacionamento linear entre as variáveis X e Y que quando dividido pelo produto dos desvios padrões das variáveis resulta na medida da correlação entre elas. A correlação entre duas variáveis pode variar entre -1 e 1 de modo que quando duas variáveis são independentes, a correlação entre elas é nula e quando o valor da correlação se aproxima de -1 ou de +1 a correlação entre as variáveis é mais forte, seja esta positiva ou negativa.

Dentro os métodos de análise de correlação mais comuns existem os métodos de correlação paramétricos, como o Método de Pearson, e os métodos de correlação não paramétricos, como os métodos de Spearman e de Kendall. Para o projeto em desenvolvimento será utilizado o método de correlação paramétrico mais comum, o Método de Pearson. tendo em vista a possibilidade de se calcular os parâmetros da média e do desvio

padrão para as variáveis a serem utilizadas, que são fatores necessários para a aplicação de um método de correlação paramétrico. Além disso, os métodos de correlação paramétricos são mais eficientes em definir a correlação entre duas variáveis justamente por contarem com mais informações sobre a distribuição destas para seu cálculo.

2.6.1 Matriz de Correlação de Pearson

Para um conjunto de dados de tamanho n, a correlação de Pearson pode ser expressa pela seguinte fórmula:

$$correlação(X,Y) = \frac{\sum_{i=1}^{n} x_i \, y_i - n \times (média(x)) \times (média(Y))}{\sqrt{(\sum_{i=1}^{n} x_i^2 - n \times média(X)^2) \times (\sum_{i=1}^{n} y_i^2 - n \times média(Y)^2)}}$$

Para um dataset com N variáveis, é possível construir uma matriz NxN em que cada célula $x_{i,j}$ representa a correlação de Pearson entre as variáveis i e j. Através desse método é possível identificar neste dataset as variáveis com maior correlação entre si e, no caso do projeto em desenvolvimento, será possível identificar as variáveis com maior correlação com o Perfil e Compra dos Clientes.

3 METODOLOGIA

Nesta seção serão introduzidas as etapas da pesquisa desenvolvida, de modo que serão analisadas as contribuições de cada uma delas para o objetivo do projeto. Na sequência, serão também apresentados os dados que serão utilizados para o projeto e o método de extração pertinente a estes.

3.1 Etapas da pesquisa

A pesquisa desenvolvida será dividida em oito macro etapas sequenciais de trabalho. Estas etapas serão desenvolvidas de forma sequencial, porém iterativa ao longo dos meses de projeto, ou seja, as etapas apresentam grande dependência uma das anteriores, mas serão constantemente feitos ajustes nas etapas prévias mediante aos aprendizados adquiridos ao longo do projeto, conforme defendido pelos princípios de melhoria contínua.

Inicialmente serão dedicados esforços para detalhar o projeto junto ao cliente, de modo a compreender os indicadores relevantes para o trabalho e o impacto necessário nestes. Para tanto, serão identificados os Objetivos e Resultados-Chave da organização nos últimos quarters para identificar os indicadores foco do projeto. O espaço amostral do projeto também será definido nesta etapa, de modo que serão utilizados dados de clientes para a análise que estejam de acordo com todas as partes envolvidas. Além disso, nesta etapa serão identificados quais são os indicadores de maior relevância para a classificação do perfil de compra de um cliente com base em pesquisas realizadas com os *stakeholders* mais importantes. Serão definidos também os períodos com relação aos quais serão extraídas as informações para treinamento e para aplicação do modelo de clusterização desenvolvido. Por fim, será importante definir o impacto esperado do projeto junto aos *stakeholders* e o intervalo para iterar as entregas do projeto. O *Business Owner* e as lideranças do time de vendas serão consultados a fim de coletar insumos a respeito da visão destes sobre todos os aspectos a serem definidos nesta fase.

Na segunda etapa do projeto serão definidos os perfis de clientes existentes atualmente na organização com relação ao comportamento de compra desses. As bases de dados que contenham informações relevantes ao projeto serão tratadas para que essas sejam facilmente

utilizadas nas etapas seguintes. Será utilizado o Databricks para o acesso aos bancos de dados e às tabelas contidas nestes. Tendo sido acessadas as tabelas, será utilizado o *PyCharm* para rodar os comandos da biblioteca *Pyspark* do *Python* para que sejam tratadas as informações selecionando apenas as tabelas que contenham dados pertinentes, garantindo a consistência do formato dos dados e a completude de informações necessárias. Uma vez desenvolvido o código para a obtenção dos datasets do projeto, este código será formatado de acordo com os princípios do manual PEP-8 para garantir a padronização das entregas de acordo com as normas da organização.

Para a definição dos perfis de clientes serão utilizados os indicadores de compra dos clientes levantados na primeira etapa e as bases de dados compiladas. Com essas informações, será utilizado o método de clusterização *K-Means* para separar os clientes em grupos distintos denominados clusters. Cada um desses clusters será analisado a fim de se compreender as performances dos clientes contidos nestes para a caracterização desses grupos.

Na terceira etapa, os perfis de clientes serão caracterizados. Isto é, serão definidos, dentre os clusters elaborados, quais são considerados bons, médios ou ruins para a organização. Além disso, estes grupos serão nomeados de forma que a classificação resultante do método *K-Means* seja mais facilmente compreendida pelos colaboradores da empresa.

A quarta etapa do projeto pressupõe a projeção do potencial máximo de compra de cada cliente, estimando-se a melhor performance possível para estes clientes. Esta projeção será feita de modo que o Potencial Máximo de Compra do cliente será a melhor performance que cada cliente atingiu no período de análise para os *KPIs* relevantes ao projeto. Será construído também um Painel de Controle que permita aos vendedores acompanharem no decorrer do mês a performance dos clientes da sua carteira de clientes no mês corrente de forma comparativa com o Potencial Máximo destes clientes. Com base nessa análise espera-se identificar mês a mês os clientes com maior diferença entre a performance no mês corrente e a performance máxima do cliente, facilitando o foco do time de vendas em priorizar suas ações tendo em vista a obtenção de melhores resultados nos *OKRs* da empresa.

Na quinta etapa será utilizado o Solver do Excel para a elaboração de um otimizador que seja capaz de definir a distribuição ideal da base de clientes da empresa entre os diferentes perfis de compra existentes para maximizar a rentabilidade do modelo de negócios. Tendo em posse a clusterização realizada e os potenciais máximos de cada cliente, será

calculado o potencial médio de um cliente que pertence a cada cluster. Adotando projeções realistas de redução ou aumento de clusters que sejam factíveis como restrições do otimizador, será definido uma porcentagem ideal de clientes pertencentes a cada perfil de compra que comporá a denominada base de clientes ideal.

A sexta etapa do projeto consistirá na análise das transições dos clientes entre clusters mês a mês. Para isso, o modelo *K-Means* desenvolvido será aplicado nos clientes mês a mês utilizando a base histórica de modo que cada cliente passará a ser classificado como pertencente a determinado cluster de acordo com seu perfil de compra no mês em questão. Na sequência serão analisadas as mudanças de cada clusters no período analisado, buscando-se compreender quais clusters aumentaram ou diminuíram no período e, principalmente, como os clientes foram influenciados a transitar entre os clusters. Entender esse processo de transição será a chave para a definição de planos de ação capazes de alterar a base de clientes atual para o formato de base de clientes ideal.

A sétima etapa consistirá na análise da correlação de cara atributo do cliente com a classificação dele dentre os perfis de compra levantados para identificar quais atributos possuem maior influência nesta classificação. Serão levantados atributos geográficos do cliente a partir de uma base de dados geográficos chamada Geodata que contém informações que vão desde a densidade demográfica da região até o Índice de Desenvolvimento Humano desta, além de outras cerca de 10 variáveis. Serão considerados também dados levantados pelo time de vendas a partir de coleta manual a respeito de atributos físicos das lojas dos clientes, como o tamanho da loja, a quantidade de caixas registradoras e a quantidade de pontos extra de venda. Ainda, serão analisadas as variáveis comportamentais dos clientes, identificando outros comportamentos de compra que possam ser relevantes para o entendimento do perfil de compra deste. A partir dessas informações será construída uma Matriz de Correlação em que será identificada a correlação entre cada uma dessas variáveis e os clusters de clientes, identificando assim quais as mais significativas para a classificação em clusters. Tendo em mãos as variáveis de maior correlação com os clusters, será identificada, para cada cluster, uma lista de variáveis que possuem maior influência na atribuição do cliente a este respectivo cluster.

Por fim, a oitava etapa do projeto envolverá a criação de planos de ação para o direcionamento da empresa quanto aos entregáveis das etapas anteriores. As apresentações de cada uma das etapas de quatro a cinco para os *stakeholders* envolvidos serão realizadas

tomando por base os conceitos do Princípio da Pirâmide para melhorar o entendimento de forma objetiva dos entregáveis. E serão utilizados os princípios do método *Lean Startup* Enxuta para direcionar planos de ação iterativos que se baseiam na clusterização de clientes, na identificação dos atributos de maior relevância para a alocação dos clientes em cada cluster e na projeção de potencial de compra dos clientes.

3.2 Coleta de dados

Como mencionado anteriormente, os dados para o projeto serão coletados a partir dos bancos de dados existentes no Databricks. Esses bancos de dados compreendem tabelas com informações provenientes de diversas aplicações de TI como o E-Commerce, o *ERP*, o *WMS*, o *CRM* e o Roteirizador.

A partir desses bancos de dados serão utilizadas quatro principais tabelas que contém dados relevantes para o projeto desenvolvido: *LINE_ITEMS, CONTRIBUTION_MARGIN_1_PER_LINE_ITEM, GEODATA E PIPEDRIVE_ORGANIZATIONS*.

A *LINE_ITEMS* é a tabela que armazena as informações pertinentes aos pedidos da empresa, com todas as quantidades vendidas de cada produto em cada pedido para cada cliente, bem como os preços e custos respectivos. As principais informações contidas nesta tabela que serão utilizadas para o projeto em questão estão expostas na Tabela 4.

Tabela 4: Schema da Tabela *LINE_ITEMS*

Nome da Coluna	Formato do Dado
order_date	DateType
order_id	StringType
order_status	StringType
store_id	StringType
store_name	StringType
customer_store_id	StringType
customer_store_name	StringType
product_id	StringType
product_name	StringType
quantity_ordered_un	DoubleType
price_un	DoubleType
revenue_ordered	DoubleType
gross_revenue	DoubleType
gross_revenue_taxes	DecimalType(38,10)
net_revenue	DoubleType
gross_cost	DoubleType
net_cost	DoubleType
gross_profit	DoubleType

Fonte: Databricks

A tabela *CONTRIBUTION_MARGIN_1_PER_LINE_ITEM* é a tabela que representa a estrutura de custos do P&L da empresa até o nível de Margem de Contribuição. O schema desta tabela pode ser visto na Tabela 5.

Tabela 5: Schema da Tabela *CONTRIBUTION_MARGIN_1_PER_LINE_ITEM*

Nome da Coluna	Formato do Dado
order_id	StringType
customer_store_name	StringType
customer_store_id	StringType
product_id	StringType
quantity_delivered_un	DoubleType
lucro_liquido	DoubleType
lucro_bruto	DoubleType
receita_liquida	DoubleType
receita_bruta	DoubleType
debito_de_impostos	DoubleType
cogs	DoubleType
custos_operacionais	DoubleType

Fonte: Databricks

A tabela *GEO_DATA* contém informações que dizem respeito à região dos clientes, fornecendo, para cada cliente, informações referentes a diversos aspectos socioeconômicos como índice de vulnerabilidade social, índice de desenvolvimento humano e densidade demográfica. O schema dessa tabela pode ser visto na Tabela 6 abaixo.

Tabela 6: Schema da Tabela *GEO_DATA*

Nome da Coluna	Formato do Dado
customer_store_id	StringType
name	StringType
ivs	FloatType
idhm	FloatType
espvida	FloatType
renda_per_capita	FloatType
populacao	IntegerType
i_gini	FloatType
ivs_infraestrutura_urbana	FloatType
t_densidadem2	FloatType

Fonte: Databricks

Por fim, a tabela *PIPEDRIVE_ORGANIZATIONS* traz informações coletadas via *CRM* da empresa, que é o Pipedrive sobre o perfil de loja do cliente. Essas informações foram coletadas manualmente pelos vendedores responsáveis por cada cliente ao longo dos últimos anos e compreendem uma forma de entender melhor o perfil de cada loja dos clientes. O schema desta fonte de dados pode ser visto na Tabela 7.

Tabela 7: Schema da Tabela *PIPEDRIVE_ORGANIZATIONS*

Nome da Coluna	Formato do Dado
customer_store_id	StringType
checkouts_quantity	IntegerType
corridors_quantity	IntegerType
gondola_tip_quantity	IntegerType
store_size_m2	DoubleType
inventory_format	StringType
customer_store_type	StringType

Fonte: Databricks

Além das informações contidas nestas tabelas foram coletadas informações referentes à experiência do *Business Owner* da Varejo ABC e das lideranças do time de vendas para definir com maior precisão os critérios relevantes para o projeto, bem como para compreender melhor o perfil dos clientes. Essas informações foram coletadas mediante diversas reuniões realizadas ao longo do projeto e foram expostas nas seções seguintes do trabalho juntamente com a aplicação delas.

4 RESULTADOS

Durante a seção de resultados do projeto, serão executadas todas as atividades descritas na metodologia do projeto. Dentro do próprio desenvolvimento do projeto serão apresentados e discutidos os resultados obtidos em cada etapa, cabendo posteriormente uma análise geral sobre o resultado geral do projeto no intuito de compilar todos os pontos abordados a ser realizada na conclusão. Ressalta-se que este projeto se desenvolve no intuito de se atingir o objetivo do trabalho desenvolvido, que é: *"Consolidar, até o final do ano de 2022, planos de ação implementáveis que sejam capazes de alterar a distribuição dos perfis de cliente da base atual da Varejo ABC para uma composição que otimize a adesão desta base à proposta de valor da empresa"*.

4.1 Detalhamento do Projeto Desenvolvido

Para dar início ao detalhamento do projeto desenvolvido, foram identificados e analisados os *OKRs* da empresa para os trimestres em que o projeto foi desenvolvido. Não sendo prudente entrar nas minúcias de quais eram exatamente os *OKRs* da organização para preservar a estratégia desta, compete ressaltar que os objetivos da organização no período estavam voltados essencialmente para a melhoria da margem de empresa.

Na sequência, utilizou-se do entendimento dos objetivos estratégicos da empresa para realizar a definição dos *KPIs* (Indicadores de Performance Chave) que possuem maior relevância na adesão de um cliente ao modelo de negócios da Varejo ABC. Para esta definição, foram consultados os especialistas de negócio em uma reunião com a presença do Business Owner e de três integrantes da liderança de vendas.

Nesta reunião, os presentes foram indagados a respeito de quais seriam os indicadores de maior interesse deles para a definição de um bom cliente para a empresa. Esse levantamento se faz necessário pois será a base do modelo de clusterização construído na sequência, haja vista que o modelo será construído a partir de um dataset que contenha os valores destes *KPIs* para cada cliente.

A primeira etapa da reunião foi a realização de um brainstorming, no qual foram levantados 12 *KPIs* relevantes para a organização que são, nominalmente: margem bruta percentual, margem bruta, receita, variedade de *SKUs*, ticket médio, variedade de categorias, sensibilidade a preço, frequência de pedidos, share de compras a preço de oferta vs a preço

regular, autonomia de compra no site, inadimplência e percentual médio de limite de crédito utilizado. Vale ressaltar que esta lista de indicadores levantados no brainstorming já consiste apenas de indicadores que são acessíveis pelos dados contidos nos bancos da empresa, sem serem necessárias coletas manuais ou levantamentos extra de dados para a obtenção de nenhum destes.

Em uma segunda etapa, foi realizado um filtro nesses indicadores de modo a restarem apenas indicadores MECE (mutuamente excludentes e coletivamente exaustivos), ou seja, nenhum indicador que traga uma informação duplicada com relação a outro indicador e que não falta nenhuma informação relevante para a definição da adesão do cliente ao modelo de negócios da empresa que delimitará seu perfil de compra. Dessa forma, foram selecionados os indicadores margem bruta, percentual, receita bruta e variedade de skus mensais como sendo os três indicadores a serem considerados durante o projeto.

O quórum da reunião deliberou também acerca do período de apuração destes indicadores. O período selecionado para a apuração foi entre junho de 2022 e outubro de 2022, período com estrutura mais similar à atual das operações da empresa devido a algumas mudanças operacionais mais significativas ocorridas em maio de 2022. Ficou então determinado que o modelo de clusterização seria treinado utilizando-se todo o período entre junho e outubro e que o modelo desenvolvido seria aplicado para clusterizar os clientes mês a mês tanto neste período quanto nos meses futuros. Ainda, buscou-se atenuar as variações de compra entre os meses calculando os indicadores como média dos meses analisados. Com esta definição, o autor já detinha de base suficientemente estável para a construção da clusterização de clientes com base no perfil de compra.

Em uma segunda reunião foi definido também o espaço amostral de clientes a ser utilizado para o projeto. Neste quórum delimitou-se como regra para seleção que todos os clientes selecionados tivessem realizado ao menos um pedido no período de apuração. Neste fórum, foi selecionada uma base de cerca de 1400 clientes disponíveis para análise.

Na etapa final de detalhamento de projeto junto ao Business Owner da organização foi delimitada como noção de sucesso para o projeto que este fosse capaz de gerar planos de ação que dessem origem a iniciativas com resultado positivo comprovado para a margem de contribuição da empresa. Também foi definido que o projeto possuiria iterações quinzenais, isto é, quinzenalmente o autor do projeto e o *Business Owner* se reuniram para executar alguns dos princípios iterativos descritos nos métodos tradicionais de melhoria contínua e no método Lean Startup, fazendo com que o projeto fosse desenvolvido em constante aprimoração a partir dos aprendizados adquiridos ao longo do curso do desenvolvimento.

4.2 Clusterização dos Clientes por Perfil de Compra

Nesta segunda etapa, os clientes foram distribuídos em clusters com base nos três indicadores mais relevantes para a determinação do perfil de compra dos clientes levantados nas etapas anteriores. Foram importadas as bibliotecas *Python* necessárias para a execução dos códigos, foi construído e analisado o dataset a ser utilizado, foram padronizadas as informações contidas neste dataset e foram definidos os parâmetros significativos ao modelo - o valor de clusters K, o método de cálculo da distância entre os pontos e o método de normalização dos dados. Com base nisso o modelo *K-Means* foi treinado e aplicado ao dataset para obter-se os resultados esperados na clusterização, atribuindo cada um dos clientes da empresa a um clusters específico. Tendo feita a clusterização, foi analisada a distribuição dos clientes com relação a cada um dos *KPIs* para melhor compreensão das performances médias de cada cluster, bem como os desvios padrões atrelados a estas médias.

Um ponto que vale ressaltar é que durante a elaboração deste trabalho, todos os dados dos clientes foram normalizados para evitar a exposição da performance efetiva de cada cliente, que configura-se como um conteúdo sensível e estratégico para a empresa. Os nomes dos clientes também foram preservados para assegurar o anonimato destes tendo sido divulgados apenas os identificadores únicos deles.

4.2.1 Importação das Bibliotecas Necessárias

A importação das bibliotecas necessárias foi dividida em diferentes categorias para facilitar a compreensão do leitor (Apêndice A). Na primeira seção foram importadas as tabelas e as bibliotecas referentes ao tratamento das informações contidas nestas tabelas. Na sequência foram importadas bibliotecas estruturais clássicas do python como a numpy, o pandas e o pyspark que servirão como ferramentas para o tratamento dos dados. Para a visualização dos gráficos que materializarão a compreensão das etapas realizadas foram importadas três bibliotecas visuais que são a *Matplotlib*, a *Seaborn* e a *Axes3D*.

Durante a aplicação efetiva do método *K-Means* foram necessárias bibliotecas para normalizar os dados dos clientes e foram testadas duas diferentes modalidades que foram importadas a partir da biblioteca de machine learning do *Pyspark*, o método *MinMaxScaler*, o métodos *RobustScaler* e o método *StandardScaler*. Por fim, foram importadas as bibliotecas

pertinentes à aplicação do modelo *K-Means* propriamente dito na última seção de importações do notebook utilizado para este projeto. Foram também definidas constantes a serem usadas durante o projeto, que são as datas iniciais e finais do período de análise, os parâmetros do Databricks que direcionam o notebook a gravar o modelo no local destinado a ele dentro da estrutura da organização e os clientes e lojas de exceção para a análise por representarem comprar outliers de falhas operacionais ou testes realizados internamente na empresa que poderiam enviesar a análise.

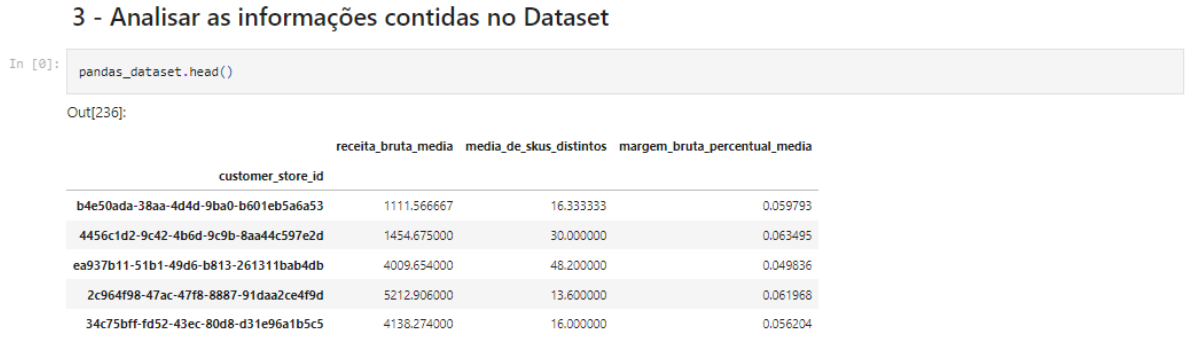
4.2.2 Construção do Dataset

Partindo da tabela *CONTRIBUTION_MARGIN_1_PER_LINE_ITEM* cujo schema foi descrito nas seções acima do trabalho, foi possível obter-se os três indicadores a serem trabalhados no projeto para cada cliente. A tabela foi filtrada para conter apenas os dados do período de análise e foram contabilizadas - excluindo os dados outliers - as margens brutas percentuais, as receitas brutas e as médias de skus distintos para cada cliente no período por meio da função *get_customers_kpis_dataset* (Apêndice A). Essa função ainda retorna ao usuário a opção de receber o dataset construído no formato Pyspark, que é o adequado para os tratamentos futuros e para a aplicação do método de clusterização, ou no formato Pandas, que é melhor para a construção das visualizações gráficas a serem construídas para análise dos dados contidos no dataset.

4.2.3 Análise das Informações do Dataset

Com o dataset construído pela função get_customers_kpis_dataset compete uma análise das informações contidas nele para garantir a consistência dessas informações para a aplicação das etapas seguintes. Assim, foi possível ter uma visualização inicial do dataset utilizando o método head() do pandas (Ilustração 2) que mostra os primeiros registros contidos no dataset. Podemos ver o resultado como sendo um data frame com o customer_store_id como index e três colunas, sendo elas receita_bruta_media, media_de_skus_distintos e margem_bruta_percentual_media.

Ilustração 2: Visualização do dataset utilizado para a clusterização



3 - Analisar as informações contidas no Dataset

In [0]:
```
pandas_dataset.head()
```

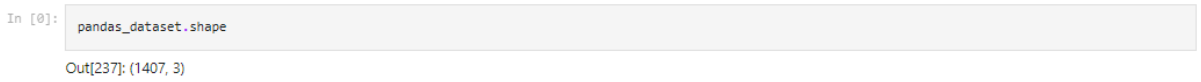
Out[236]:

customer_store_id	receita_bruta_media	media_de_skus_distintos	margem_bruta_percentual_media
b4e50ada-38aa-4d4d-9ba0-b601eb5a6a53	1111.566667	16.333333	0.059793
4456c1d2-9c42-4b6d-9c9b-8aa44c597e2d	1454.675000	30.000000	0.063495
ea937b11-51b1-49d6-b813-261311bab4db	4009.654000	48.200000	0.049836
2c964f98-47ac-47f8-8887-91daa2ce4f9d	5212.906000	13.600000	0.061968
34c75bff-fd52-43ec-80d8-d31e96a1b5c5	4138.274000	16.000000	0.056204

Fonte: Elaboração Própria

Para analisar o espaço amostral disponível no dataset, podemos utilizar o método shape do pandas (Ilustração 3). Nesse método é retornado como output o tamanho do dataset, que é de 1407 linhas por 3 colunas, o que indica a utilização de 1407 clientes disponíveis para análise no espaço amostral a ser utilizado para o projeto e de 3 variáveis contabilizadas para cada um destes clientes.

Ilustração 3: Formato do dataset utilizado para clusterização



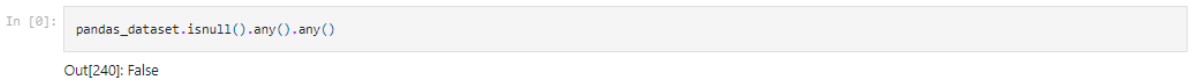
In [0]:
```
pandas_dataset.shape
```

Out[237]: (1407, 3)

Fonte: Elaboração Própria

Tendo definido o formato do dataset, serão feitas duas verificações finais para garantir sua consistência, será verificado se existe algum registro nulo no dataset (Ilustração 4) e se todos os valores contidos nele são reais (Ilustração 5). Ambas as características são necessárias para que as informações utilizadas durante a clusterização sejam consistentes com os métodos a serem aplicados pelas bibliotecas que aplicam o método K-Means no dataset.

Ilustração 4: Verificação da existência de valores nulos no dataset



In [0]:
```
pandas_dataset.isnull().any().any()
```

Out[240]: False

Fonte: Elaboração Própria

Ilustração 5: Verificação de valores irracionais no dataset

```
In [0]: pandas_dataset.applymap(np.isreal)
```

Out[241]:

customer_store_id	receita_bruta_media	media_de_skus_distintos	margem_bruta_percentual_media
b4e50ada-38aa-4d4d-9ba0-b601eb5a6a53	True	True	True
4456c1d2-9c42-4b6d-9c9b-8aa44c597e2d	True	True	True
ea937b11-51b1-49d6-b813-261311bab4db	True	True	True
2c964f98-47ac-47f8-8887-91daa2ce4f9d	True	True	True
34c75bff-fd52-43ec-80d8-d31e96a1b5c5	True	True	True
...	...	...	...
17d7693f-5eb9-4fcd-a2fb-3ceeacc1f4c4	True	True	True
d13acaf1-fa7e-4f6f-9b75-5936a6f6bc72	True	True	True
ba7b2605-588d-4f3d-bb54-e07bf67f44d1	True	True	True
85cdd0b8-2d97-4f32-83e1-16e6cfb53d4b	True	True	True
8e2ba4ef-ee98-4837-be94-912681a307df	True	True	True

Fonte: Elaboração Própria

Como podemos observar nas imagens abaixo, o dataset passou em ambas as validações e se demonstra consistente para análise.

4.2.4 Padronização das Informações no Dataset

O dataset consistente construído e verificado nas etapas anteriores será sujeito a dois métodos de padronização necessários para a correta execução da biblioteca *K-Means* do *Pyspark*. O primeiro desses métodos consiste na compilação das três colunas que compreendem as variáveis de interesse para cada cliente em uma única coluna vetorizada através do método *VectorAssembler* e o segundo consiste na normalização deste vetor utilizando algum dos métodos de standardization da biblioteca de machine learning do Pyspark.

Através da função *get_assembled_pyspark_dataset* (Apêndice A) o dataset obtido pela função a função *get_customers_kpis_dataset* será sujeito ao método *Vector Assembler* da biblioteca *pyspark.ml.feature* e gerará um vetor que compila as colunas passadas no parâmetro inputColumns em uma coluna nova a ser criada com o nome fornecido no parâmetro outputCol.

O dataset resultante após a padronização em um formato de vetor pode ser observado na Ilustração 6 e retrata o dataset anterior acrescido de uma nova coluna nomeada features, que representa o conjunto das variáveis a serem utilizadas no modelo em um único vetor.

Ilustração 6: Resultado da função *get_assembled_pyspark_dataset*

```
In [0]: assembled_spark_dataset = get_assembled_pyspark_dataset()
```

```
In [0]: display(assembled_spark_dataset.limit(10))
```

customer_store_id	receita_bruta_media	media_de_skus_distintos	margem_bruta_percentual_media	features
b4e50ada-38aa-4d4d-9ba0-b601eb5a6a53	1111.5666666666668	16.333333333333332	0.05979322067952141	Map(vectorType -> dense, length -> 3, values -> List(1111.5666666666668, 16.333333333333332, 0.05979322067952141))
4456c1d2-9c42-4b6d-9c9b-8aa44c597e2d	1454.675	30.0	0.06349526430302302	Map(vectorType -> dense, length -> 3, values -> List(1454.675, 30.0, 0.06349526430302302))
ea937b11-51b1-49d6-b813-261311bab4db	4009.6539999999995	48.2	0.04983559092629939	Map(vectorType -> dense, length -> 3, values -> List(4009.6539999999995, 48.2, 0.04983559092629939))
2c964f98-47ac-47f8-8887-91daa2ce4f9d	5212.906	13.6	0.06196822041295201	Map(vectorType -> dense, length -> 3, values -> List(5212.906, 13.6, 0.06196822041295201))
34c75bff-fd52-43ec-80d8-d31e96a1b5c5	4138.274	16.0	0.05620407159119963	Map(vectorType -> dense, length -> 3, values -> List(4138.274, 16.0, 0.05620407159119963))
c938328f-8398-4771-812f-f1970fb1c316	248.76666666666665	1.6666666666666667	0.06711502076912763	Map(vectorType -> dense, length -> 3, values -> List(248.76666666666665, 1.6666666666666667, 0.06711502076912763))

Fonte: Elaboração Própria

Na sequência, o dataset será exposto a um método de normalização que transformará a coluna features em uma coluna denominada features_scaled. Esse processo será feito por meio da função *get_standardized_pyspark_dataset* (Apêndice A) que aceita três métodos de normalização que são os mais comumente utilizados no âmbito de *Machine Learning*, que são o *MinMaxScaler*, o *StandardScaler* e o *RobustScaler*.

Assim como feito para o dataset resultante da aplicação do *VectorAssembler*, será exposto na Ilustração 7 um exemplo das primeiras colunas resultantes deste método de modo que seja possível observar a nova coluna criada denominada *features_scaled*.

Ilustração 7: Resultado da função *get_standardized_pyspark_dataset*

```
In [0]: standardized_spark_dataset = get_standardized_pyspark_dataset('RobustScaler')
```

```
In [0]: display(standardized_spark_dataset.limit(10))
```

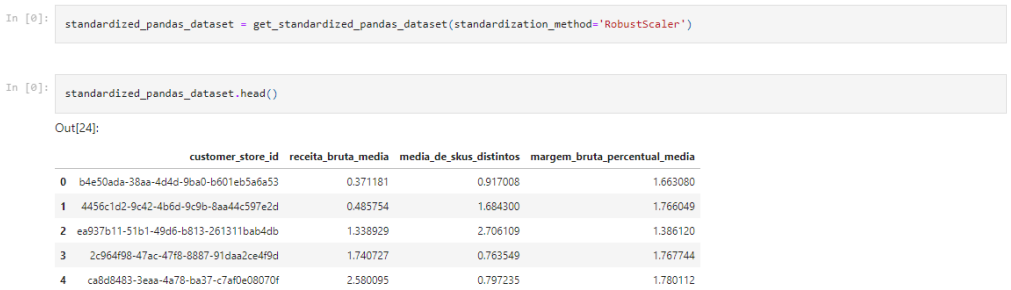
customer_store_id	receita_bruta_media	media_de_skus_distintos	margem_bruta_percentual_media	features	features_scaled
b4e50ada-38aa-4d4d-9ba0-b601eb5a6a53	1111.5666666666666	16.333333333333332	0.05979322067952141	Map(vectorType -> dense, length -> 3, values -> List(1111.5666666666666, 16.333333333333332, 0.05979322067952141))	Map(vectorType -> dense, length -> 3, values -> List(0.3711814129844826, 0.9170078163695514, 1.6630804792675506))
4456c1d2-9c42-4b6d-9c9b-8aa44c597e2d	1454.675	30.0	0.06349526430302302	Map(vectorType -> dense, length -> 3, values -> List(1454.675, 30.0, 0.06349526430302302))	Map(vectorType -> dense, length -> 3, values -> List(0.4857543304643916, 1.6843000708828497, 1.7660486153484203))
ea937b11-51b1-49d6-b813-261311bab4db	4009.654	48.2	0.04983559092629938	Map(vectorType -> dense, length -> 3, values -> List(4009.654, 48.2, 0.04983559092629938))	Map(vectorType -> dense, length -> 3, values -> List(1.338929172608225, 2.7061087805517787, 1.3861203243510405))
2c964f98-47ac-47f8-8887-91daa2ce4f9d	5212.906	13.6	0.06355620224113	Map(vectorType -> dense, length -> 3, values -> List(5212.906, 13.6, 0.06355620224113))	Map(vectorType -> dense, length -> 3, values -> List(1.7407267353902487, 0.7635493654668919, 1.7677435348419817))
ca8d8483-3eaa-4a78-ba37-c7af0e08070f	7726.5380000000005	14.2	0.06400087438384436	Map(vectorType -> dense, length -> 3, values -> List(7726.5380000000005, 14.2, 0.06400087438384436))	Map(vectorType -> dense, length -> 3, values -> List(2.5800947242495265, 0.7972353668845488, 1.780111585129589))

Fonte: Elaboração Própria

Ainda, é criada uma terceira função nesta seção que transforma o dataset normalizado em um dataset no formato Pandas, que será utilizado para a construção das visualizações pertinentes ao tópico de análise de distribuição dos clientes nos *KPIs*. Esse dataset é criado por meio da função *get_standardized_pandas_dataset*, que basicamente volta o vetor construído para o formato de colunas separadas e transforma o dataset através do método *toPandas*() (Apêndice A). Vale ressaltar que este não será o data frame utilizado para aplicação do método *K-Means* de clusterização, mas que será de grande importância haja vista a facilidade da biblioteca Pandas de lidar com as bibliotecas de visualização importadas no início desta etapa do projeto.

A mesma porção de dados exposta anteriormente no formato *Pyspark* pode ser exibida neste tópico no formato Pandas para visualização do resultado obtido pela função *get_standardized_pandas_dataset* (Ilustração 8).

Ilustração 8: Resultado da função get_standardized_pandas_dataset



```
In [0]: standardized_pandas_dataset = get_standardized_pandas_dataset(standardization_method='RobustScaler')

In [0]: standardized_pandas_dataset.head()
```

Out[24]:

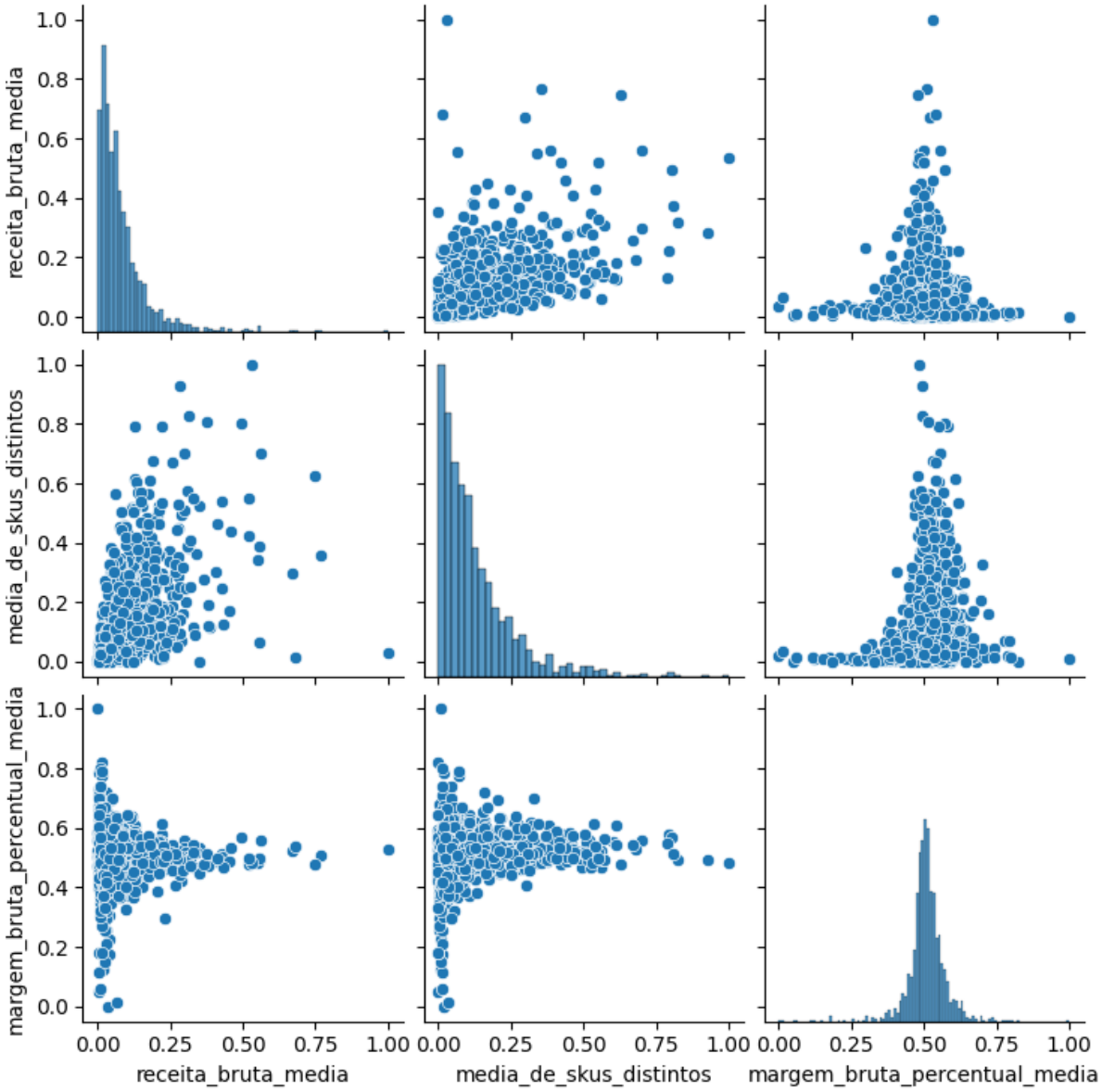
	customer_store_id	receita_bruta_media	media_de_skus_distintos	margem_bruta_percentual_media
0	b4e50ada-38aa-4d4d-9ba0-b601eb5a6a53	0.371181	0.917008	1.663080
1	4456c1d2-9c42-4b6d-9c9b-8aa44c597e2d	0.485754	1.684300	1.766049
2	ea937b11-51b1-49d6-b813-261311bab4db	1.338929	2.706109	1.386120
3	2c964f98-47ac-47f8-8887-91daa2ce4f9d	1.740727	0.763549	1.767744
4	ca8d8483-3eaa-4a78-ba37-c7af0e08070f	2.580095	0.797235	1.780112

Fonte: Elaboração Própria

4.2.5 Análise da distribuição dos clientes com base nos KPIs

A partir do dataset exibido no Gráfico 10, podemos analisar a distribuição dos clientes da Varejo ABC com base em cada uma das features (*KPIs*) que compõem este dataset. No Gráfico 10 as variáveis são comparadas duas a duas por meio da biblioteca de visualização *SeaBorn* do *Python*. Com esta visualização podem-se destacar três principais aspectos. A receita bruta e a quantidade de skus distintos que um cliente compra aparentam possuir uma correlação positiva, tendo em vista que o aumento de uma das variáveis quase sempre vem acompanhado do aumento da outra variável, embora existam vários clientes mais deslocados para um eixo ou para outro. Comparando a margem bruta com a receita bruta, é notório que em níveis muito baixos de receita bruta mensal por parte de um cliente, a margem bruta tende a sempre ser bastante baixa, o que pode ser explicado pela concentração das compras do cliente em um sortimento menor de itens buscando melhores preços em oportunidades de compra específicas. Ao analisarmos a variedade de skus mensal e a margem bruta de forma comparativa, vemos um comportamento bastante similar à observação anterior, indicando que clientes com baixa variabilidade nos skus comprados também não são os mais rentáveis para a empresa.

Gráfico 10: Distribuição dos clientes comparando KPIs 2 a 2



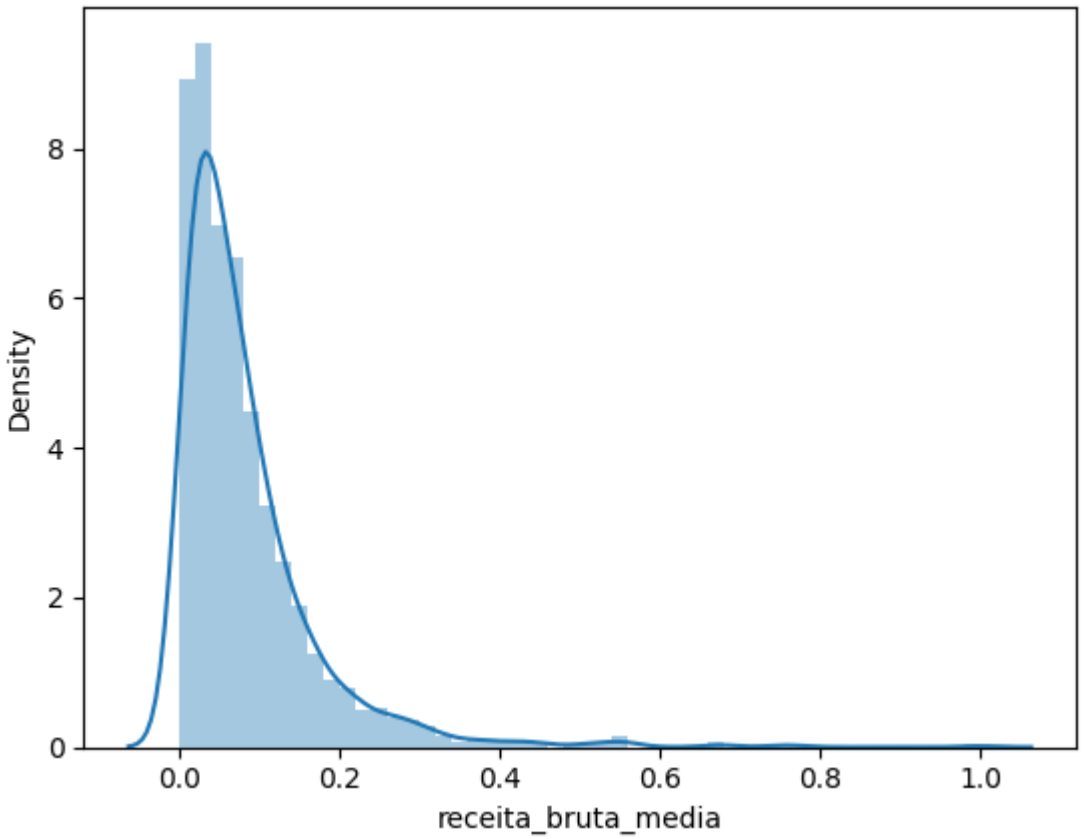
Fonte: Elaboração Própria

Para compreender melhor cada uma das features de maneira independente, pode-se construir um histograma de distribuição dos clientes com base em cada uma delas. Esses histogramas são uma visão mais detalhada dos histogramas apresentados na diagonal do Gráfico 10 e foram construídos fazendo uso do método distplot da biblioteca *SeaBorn*.

No Gráfico 11 podemos identificar o comportamento normalizado dos clientes quanto à receita bruta. Devido ao fato de todas as informações estarem normalizadas, todos os clientes possuem receita bruta distribuída entre 0 e 1. Essa distribuição possui grande similaridade com uma curva de distribuição normal e destaca que a grande quantidade de

clientes concentrados nos valores mais baixos de receita bruta e uma densidade bem menor de clientes com alta adesão a este indicador, ou seja, com elevada receita bruta mensal média.

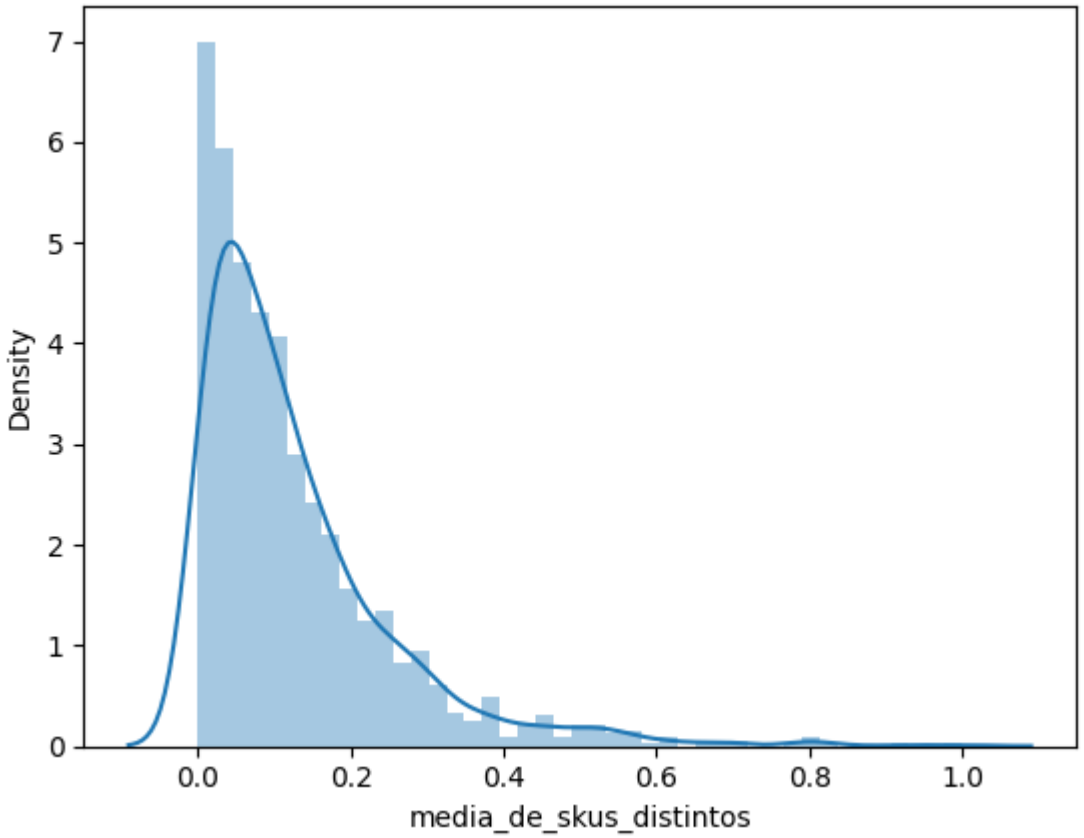
Gráfico 11: Distribuição dos clientes com relação à receita bruta



Fonte: Elaboração Própria

De maneira análoga, pode ser analisada a distribuição dos clientes quanto à variedade de skus mensais comprados por cada um deles no Gráfico 12. Neste gráfico, que também apresenta valores de 0 a 1 e apresenta distribuição aparentemente normal, podemos identificar que quanto mais longe do 0 estiver um cliente, mais completas tendem a ser as compras dele com a empresa para abastecimento de sua loja.

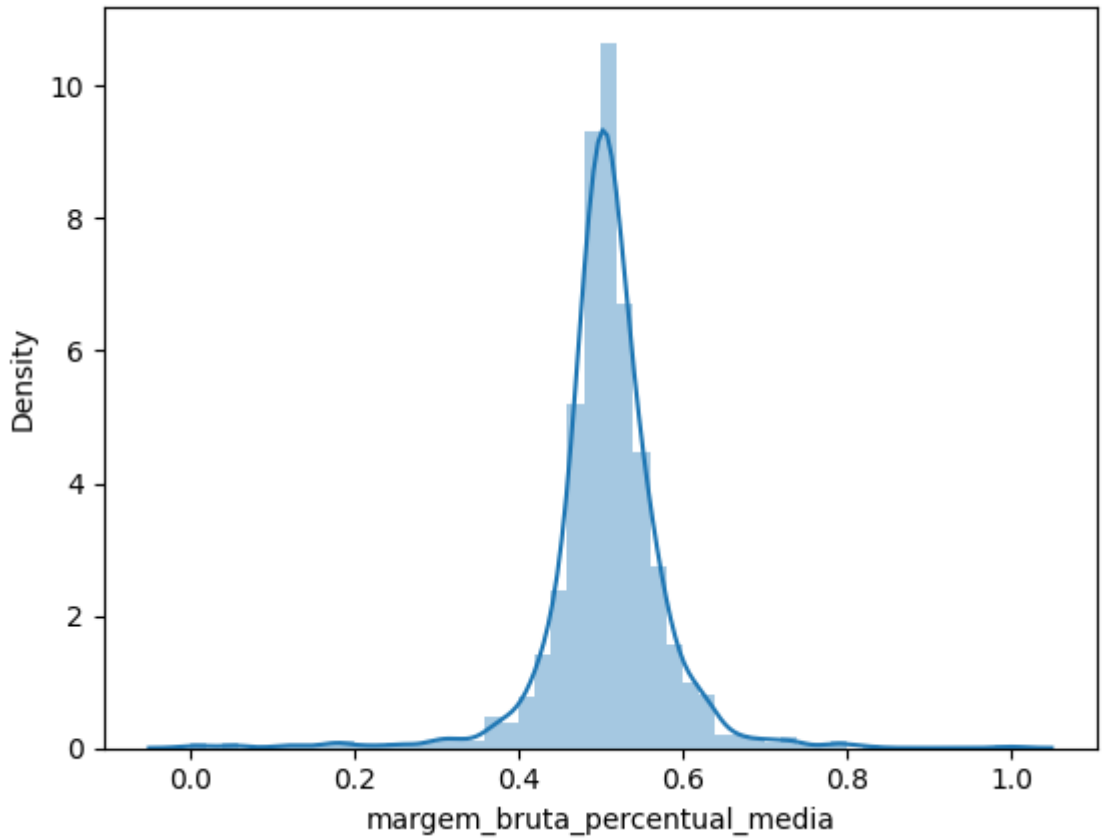
Gráfico 12: Distribuição dos clientes com relação à variedade de skus



Fonte: Elaboração Própria

Por fim, no Gráfico 13 é possível observar a distribuição dos clientes quanto à margem bruta. Neste gráfico, quanto mais próximo de 1 estiver um cliente, maior é a sua margem bruta relativa ao resto do espaço amostral analisado.

Gráfico 13: Distribuição dos clientes com relação à margem bruta



Fonte: Elaboração Própria

4.2.6 Definição do dos parâmetros do modelo K-Means

O método *K-Means* pressupõe um número K de clientes pré determinado para o treinamento do modelo. Embora o método não seja capaz de inferir um número ótimo de clusters, existem métodos capazes de auxiliar o cientista de dados a realizar esta definição. Os dois métodos mais comuns são o Método do Cotovelo e o Score da Silhueta.

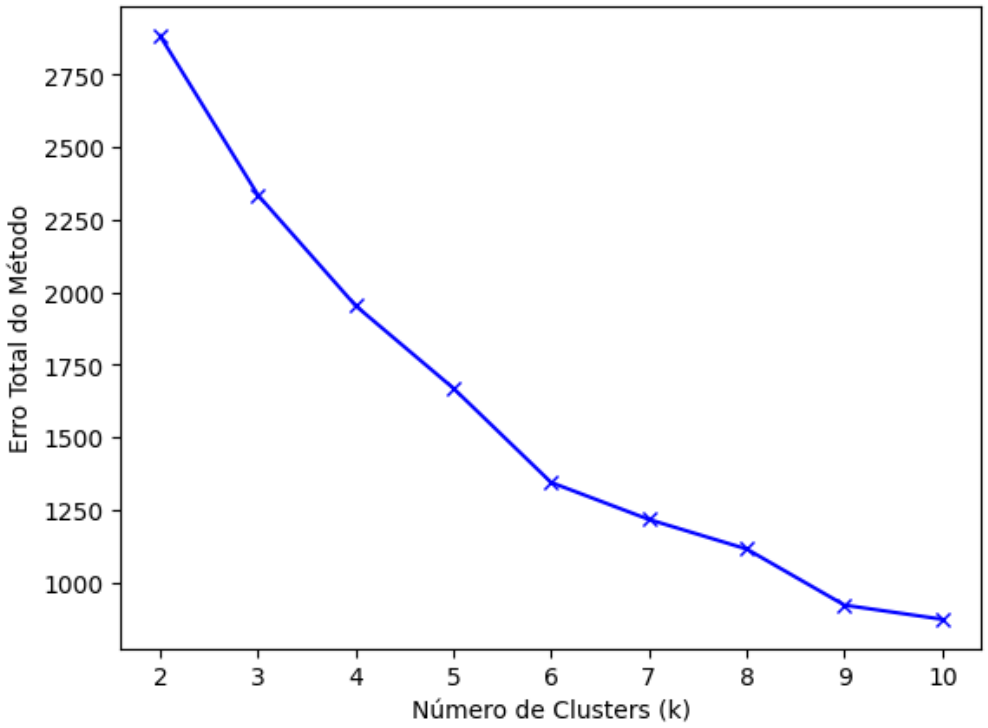
Esses dois métodos dependem de dois parâmetros que diferenciam os resultados das curvas obtidas. Primeiramente os métodos possuem resultados que variam de acordo com o método de normalização utilizado. É possível que sejam utilizados os dados brutos contidos no *pyspark_dataset* ou os dados normalizados contidos no *standardized_pyspark_dataset*, sejam estes normalizados pelo método *MinMaxScaler*, pelo método *StandardScaler* ou pelo método *RobustScaler*. Ainda, estes métodos pressupõem um método de cálculo para a distância entre os pontos para construção das curvas. Nesses métodos podem ser utilizados a distância Euclidiana, a distância de Manhattan e a distância baseada na similaridade de cossenos.

Neste projeto serão construídos tanto o Método do Cotovelo quanto o Score da Silhueta para uma análise mais completa acerca da quantidade de clusters a ser utilizada. Esses métodos serão calculados tanto para os dados normalizados quanto para os dados brutos. E quanto ao método de cálculo de distância, será utilizada a distância euclidiana que é a mais comumente utilizada na literatura para aplicação no método *K-Means*.

A função *plot_elbow_method_chart* (Apêndice A) constrói o gráfico do erro obtido no método *K-Means* para cada valor de K clusters a partir do cálculo do erro como sendo a soma das distâncias entre cada ponto e o centro do cluster ao qual ele está atribuído.

Após sucessivas iterações testando-se a performance da clusterização para cada método de normalização e para cada método de mensuração de distância entre os pontos foi observada a performance da clusterização para cada caso. As melhores performances foram adquiridas utilizando-se o método de normalização RobustScaler e o método da distância euclidiana entre os pontos. Sendo assim, as análises do Método do Cotovelo foram realizadas utilizando esses parâmetros na função *plot_elbow_method_chart* conforme resultante no Gráfico 14.

Gráfico 14: Método do Cotovelo aplicado nos dados normalizados



Fonte: Elaboração Própria

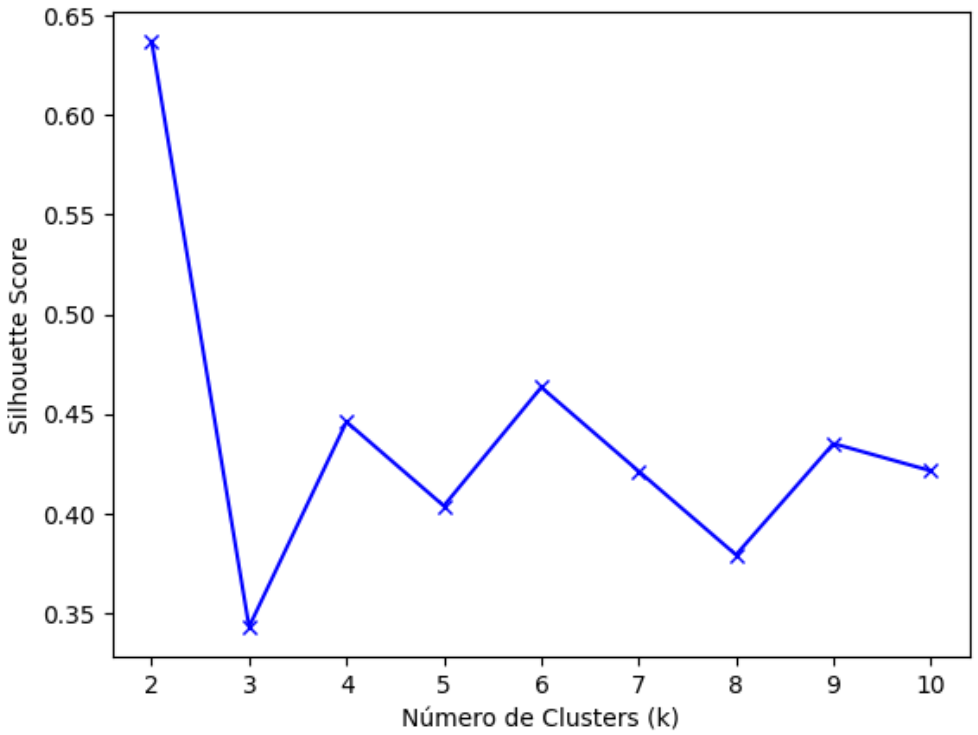
Tomando a interpretação dos gráficos conforme a revisão da literatura, sabe-se que o Método do Cotovelo direciona a aplicação do método *K-Means* para o número de clusters em que a curva passa a contar com uma diminuição menos significativa de erro para cada novo cluster adicionado, isto é, onde a curva passa a se comportar aproximadamente como uma reta de relativamente baixa derivada dando a esta um formato similar a um cotovelo, de onde vem o nome do método. Na aplicação do método para os dados normalizados, obteve-se o ponto do cotovelo para K=6.

Ressalta-se que este método não define um número determinístico e fixo de clusters que deve ser utilizado tendo em vista que ele apenas indica a partir de onde não se haverá ganhos tão mais expressivos para compensar o esforço da adição de um novo cluster, mas que, se desejável, é possível quebrar o grupo de clientes em mais ou menos clusters definidos pelo método se uma análise qualitativa da clusterização determinar relevante. Sendo assim, a aplicação deste projeto deveria possuir aproximadamente seis clusters de acordo com esse método podendo, entretanto, ter também algum número próximo de clusters como cinco ou sete.

Para complementar este método, pode-se aplicar o método Score da Silhueta construído pela função *plot_silhouette_score_chart* (Apêndice A). Neste método é usado o parâmetro evaluate que retorna um valor contido entre -1 e 1 para avaliar quão bem divididos estão os dados entre os clusters a cada valor de K testado. O valor -1 indica que os clusters estão em seu pior estado de divisão possível, 0 indica que os clusters possuem sobreposição e 1 indica que os clusters tem o melhor desempenho possível em clusterização. Assim, deve-se buscar o ponto da curva em que os dados mais se aproximem de 1 para um número de clusters que equilibre tanto uma quantidade suficiente de separações - evitando um valor de K muito baixo que não permitirá tomadas de decisão efetivas - e uma quantidade de separações que seja possível de ser manejada - evitando tantos clusters que se torna impossível tratá-los efetivamente na prática.

Assim, como no caso do Método do Cotovelo, a partir desta função serão plotadas a curva com o método de normalização *RobustScaler* e a medida euclidiana de distância entre os pontos, resultando no Gráfico 15.

Gráfico 15: Score da Silhueta aplicado nos dados normalizados

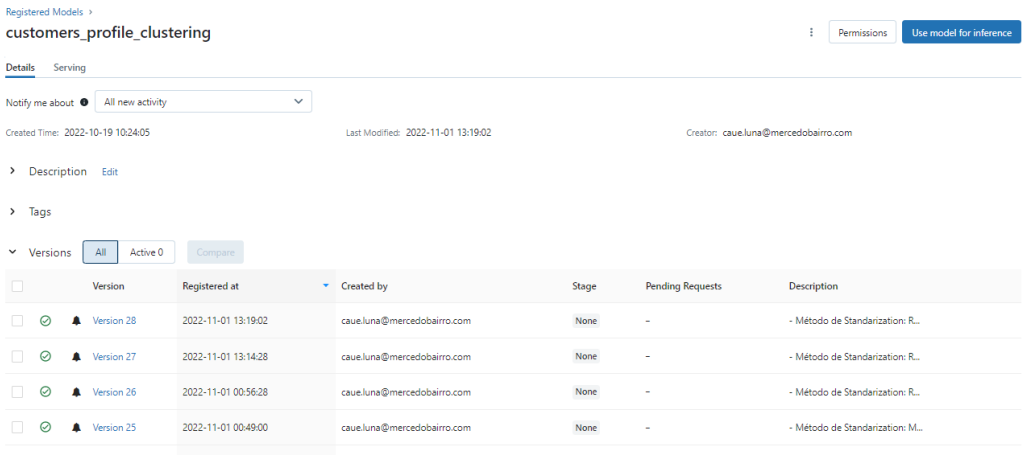


Fonte: Elaboração Própria

Ao interpretarmos os gráficos acima podemos observar que o gráfico aplicado aos dados normalizados indica K=6 como o melhor valor para a clusterização ao se descontar K=2 por ser um número de clusters tão baixo que não permitiria a definição de planos de ação concretos.

Para a definição do melhor parâmetro de normalização (*MinMaxScaler*, *StandardScaler*, *RobustScaler* ou sem normalização), do método de mensuração de distância dos pontos (distância euclidiana, distância cosseno ou distância de Manhattan) e do número de clusters utilizado (iterando-se nas proximidades do valor de K=6 definido pelo Método do Cotovelo e pelo Método da Silhueta) foram registradas 28 versões do modelo *K-Means* que foram gravados no *Databricks* (Ilustração 9) até que se chegasse a versão ideal para aplicação prática.

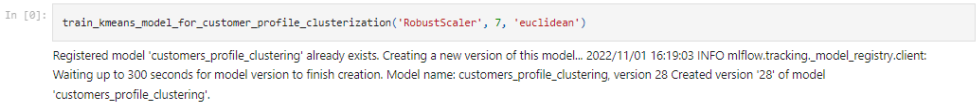
Ilustração 9: Parâmetros de cada variação do modelo K-Means



Fonte: Elaboração Própria

A versão ideal utilizando o *RobustScaler*, a distância euclidiana e 7 clusters foi definida iterativamente como sendo a de melhor performance para o modelo e foi gravada utilizando a função *train_kmeans_model_for_customer_profile_clusterization* e os parâmetros fornecidos como podemos observar na Ilustração 10.

Ilustração 10: Versão final do modelo *K-Means* treinada



Fonte: Elaboração Própria

Haja vista a iteração do modelo, a versão 28 que foi determinada como sendo a versão ótima de aplicação será utilizada nas etapas seguintes do projeto.

4.2.7 Treinar o modelo *K-Means*

Uma vez realizadas todas as definições do tópico anterior, é possível treinar um modelo *K-Means* a partir da biblioteca de *Machine Learning* do *Pyspark*. A função *train_kmeans_model_for_customer_profile_clusterization* (Apêndice A) realiza todos os processos de obtenção do *dataset*, vetorização dos dados e normalização dos dados descritos no tópico 3.3.2.4 e toma como default os parâmetros identificados no tópico 3.3.3.6.

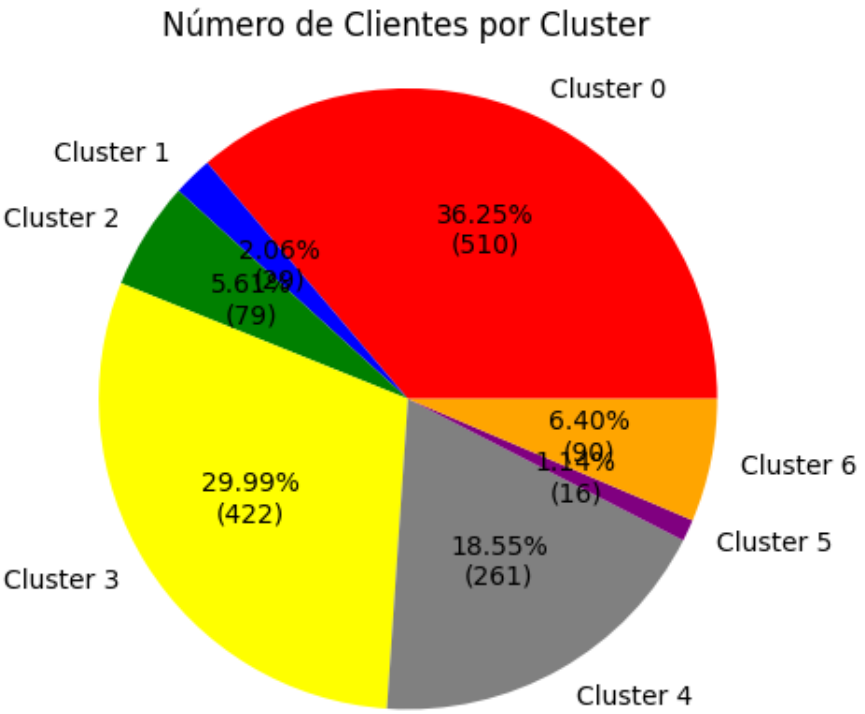
4.2.8 Aplicar o modelo *K-Means* de melhor performance treinado no dataset

Uma vez tendo as versões do modelo treinado gravadas no *Databricks*, foi possível construir uma função denominada *get_customers_clusterized_by_customer_profile* (Apêndice A) que aplica o método *K-Means* otimizado a partir da definição do parâmetro *model_version* = 28 para executar a versão otimizada do modelo.

4.2.9 Distribuição dos clientes entre os clusters

Após a aplicação do método *K-Means* é possível analisar a distribuição dos clientes entre os 7 clusters definidos através do Gráfico 16. Essa distribuição considera o perfil de compra médio dos clientes em todo o período analisado. Neste gráfico são indicados quantos clientes estão alocados em cada cluster e qual porcentagem do total do espaço amostral selecionado para o projeto esses clientes representam.

Gráfico 16: Distribuição dos clientes entre os clusters do método KMeans



Fonte: Elaboração Própria

No gráfico é possível observar que os clusters 0, 3 e 4 são os mais representativos em termos do número de clientes contidos neles. Sendo, portanto, os clusters 1, 2, 5 e 6 de menor tamanho. Em uma primeira análise superficial, esses quatro últimos clusters parecem pouco significativos perante os três clusters de maior volume, entretanto será possível identificar no tópico seguinte que estes possuem perfis de compra de extrema relevância para empresa e que fogem ao padrão médio de clientes da empresa.

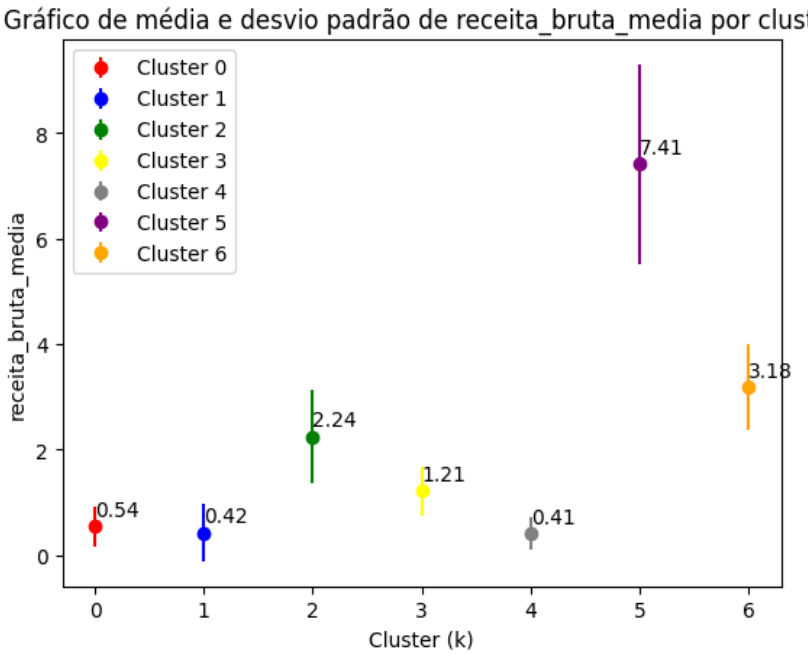
4.2.10 Análise da Média e Desvio Padrão para cluster

Nas etapas seguintes, serão realizadas quatro análises diferentes da distribuição dos clientes com relação aos *KPIs* relevantes a esta etapa do projeto. A primeira delas e mais relevante destas e a análise da média e do desvio padrão dos *KPIs* para os clientes de cada cluster. Para essa construção será estabelecido um data frame contendo as principais estatísticas de cada *KPI* para os clusters a partir da função *get_statistic_of_clusterized_pyspark_dataset* (Apêndice A).

A partir do dataset com as estatísticas de cada cluster será construída a função *plot_statistics_of_clusterized_customers_pandas_dataset* (Apêndice A) utilizando a biblioteca *matplotlib* para a construção de um gráfico a partir do dataset anterior convertido na bilioteca Pandas para melhor visualização. Essa função criará um gráfico para cada uma das features analisadas no projeto mostrando todos os clusters centralizados em cada gráfico.

O primeiro dos gráficos plotados de forma normalizada é o gráfico da receita bruta média para cada cluster (Gráfico 17). Neste gráfico podemos observar o destaque positivo para o Cluster 5, que possui patamares muito altos de receita bruta perante os demais clusters de clientes. Ainda, os Clusters 2 e 6 também possuem receita relativamente alta, embora muito menor do que a do Cluster 5. O Cluster 3 pode ser considerado como um cluster de receita média e os Clusters 0, 1 e 4 podem ser considerados Clusters de baixa receita bruta.

Gráfico 17: Média e Desvio Padrão da receita bruta dos clusters de clientes

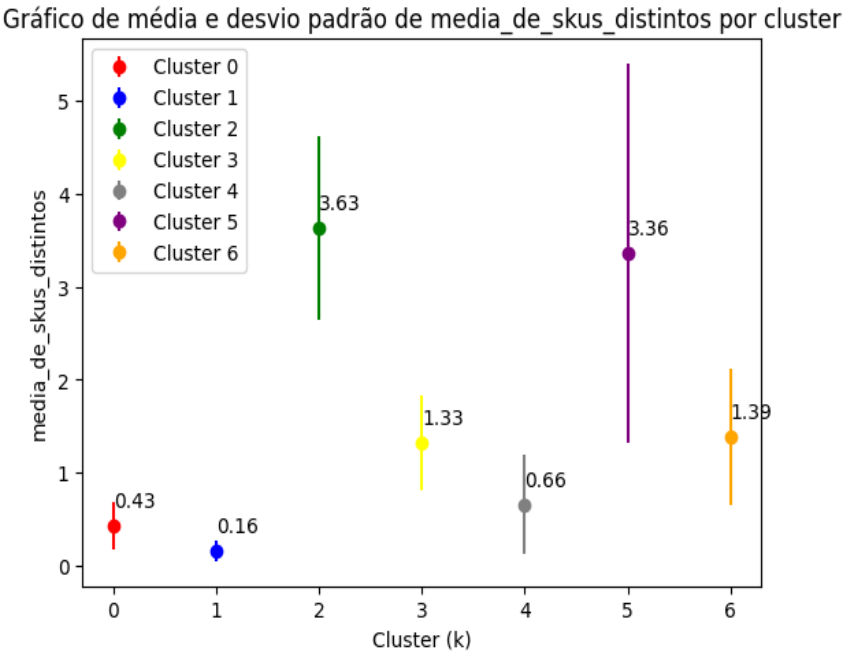


Fonte: Elaboração Própria

No Gráfico 18 é destacada a relação de cada cluster no que diz respeito à média de skus distintos mensais comprados por estes. É notório o comportamento destaque dos Clusters 2 e 5 que possuem uma variedade de produtos comprados muito maior do que todo o resto do espaço amostral utilizado. Estes, entretanto, se diferenciam pelo desvio padrão, que é muito menor entre os clientes do Cluster 2, conferindo a ele um melhor patamar geral de classificação neste KPI. Os Clusters 3 e 6 possuem performance média e os Clusters 0 e 4

possuem baixa performance. Nesse quesito, o Cluster 1 possui um destaque negativo por possuir uma performance muito menor do que os clusters supracitados.

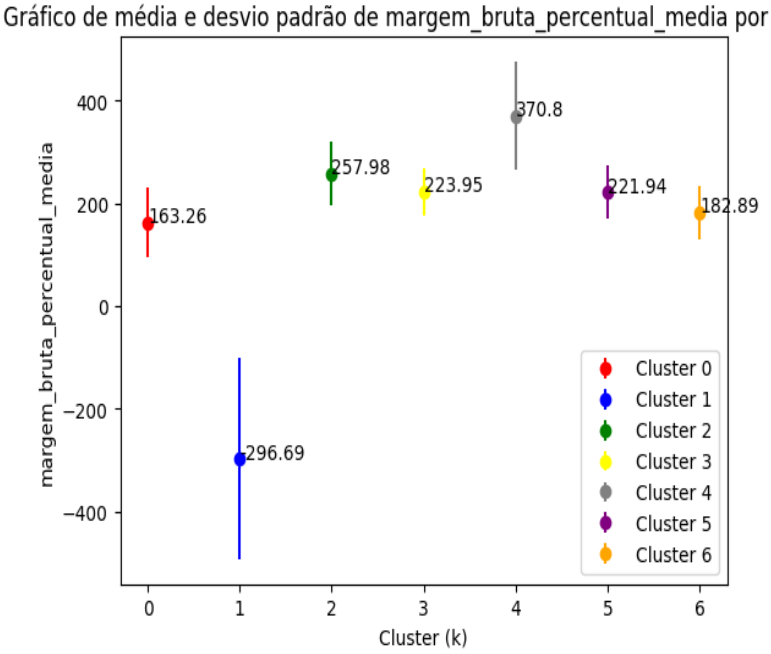
Gráfico 18: Média e Desvio Padrão de skus distintos dos clusters de clientes



Fonte: Elaboração Própria

No Gráfico 19 é possível analisar o terceiro indicador relevante para o projeto, que é a margem bruta percentual. Os postos de melhor e pior clusters ficam, respectivamente, com os Clusters 4 e 1. Ademais, os clusters 2, 3 e 5 são considerados de performance mediana neste indicador e os clusters 0 e 6 são clusters de baixa margem bruta, embora não possuam patamar tão nocivo neste indicador quanto o cluster 1.

Gráfico 19: Média e Desvio Padrão da margem bruta dos clusters de clientes



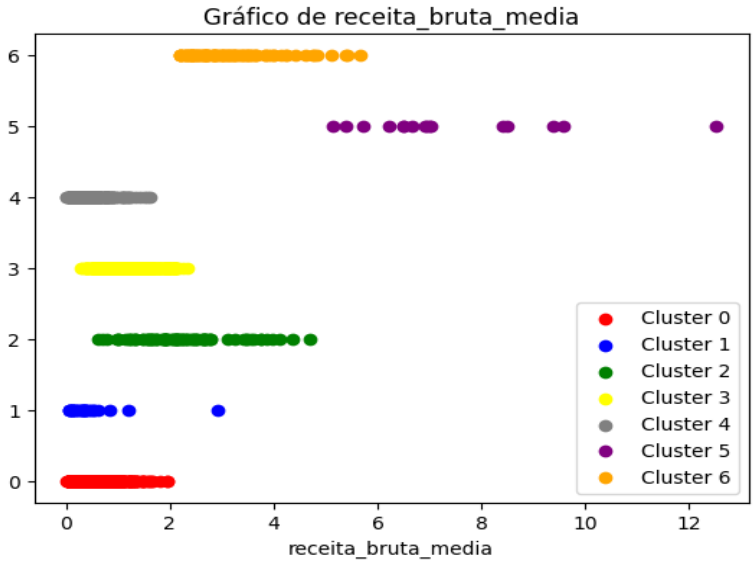
Fonte: Elaboração Própria

4.2.11 Análise 1D da distribuição dos clientes de cada cluster entre os KPIs

Para melhor compreensão do efeito do desvio padrão sobre a média da performance dos clusters em cada indicador é possível fazer uma plotagem 1D dos clientes de cada cluster em cada indicador para ver melhor como se dá a distribuição desses entorno da média do cluster ao qual pertence. No Gráfico 20 temos a distribuição da receita bruta entre os clusters de clientes, com destaque para o desvio padrão do Cluster 5, que é o mais elevado de todos. Ressalta-se que alguns dos clientes de melhor performance em receita estão neste cluster e os piores clientes deste estão próximos dos melhores clientes dos demais clusters.

Um destaque de assertividade da média de são para os clusters 0, 1, 3 e 4, que apresentam clientes bastante concentrados, ou seja, com baixo desvio padrão. O desvio padrão dos clusters 2 e 6 também possuem um desvio padrão relativamente alto, com destaque para a interseção dos valores desse indicador entre esses dois clusters, de modo que apenas por esse indicador é praticamente impossível diferenciá-los.

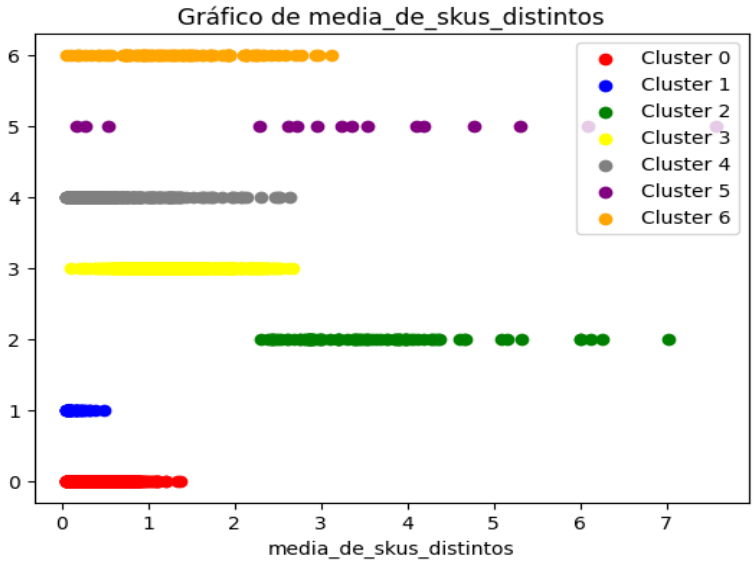
Gráfico 20: Distribuição da receita bruta entre os clusters de clientes



Fonte: Elaboração Própria

Em termos de média de skus comprados por cada cliente, o Gráfico 21 destaca também uma grande concentração dos clientes dos Clusters 0 e 1 ao redor da média destes, com um desvio padrão bastante baixo para ambos os casos. Já os clientes dos demais clusters possuem um desvio padrão relativamente alto, em especial no que se trata dos clientes dos clusters 2 e 5.

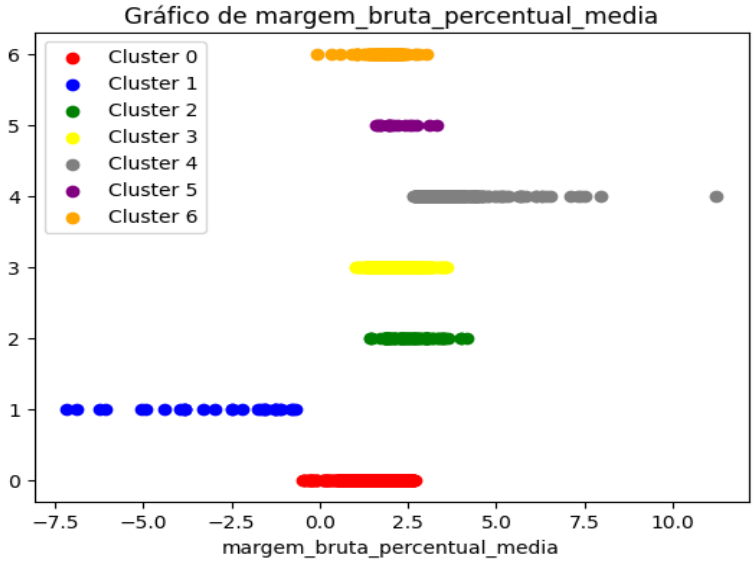
Gráfico 21: Distribuição da média de skus distintos entre os clusters de clientes



Fonte: Elaboração Própria

No caso da margem bruta, pode-se identificar no Gráfico 22 uma boa concentração dos clientes de cada cluster para cinco dos sete clusters. Porém é nítida a dispersão dos clientes dos Clusters 1 e 4. No Cluster 1, os clientes vão de valores extremamente baixos de margem bruta até valores razoáveis. E no caso do Cluster 4 os clientes vão de patamares médios até patamares extremamente altos.

Gráfico 22: Distribuição da margem bruta entre os clusters de clientes



Fonte: Elaboração Própria

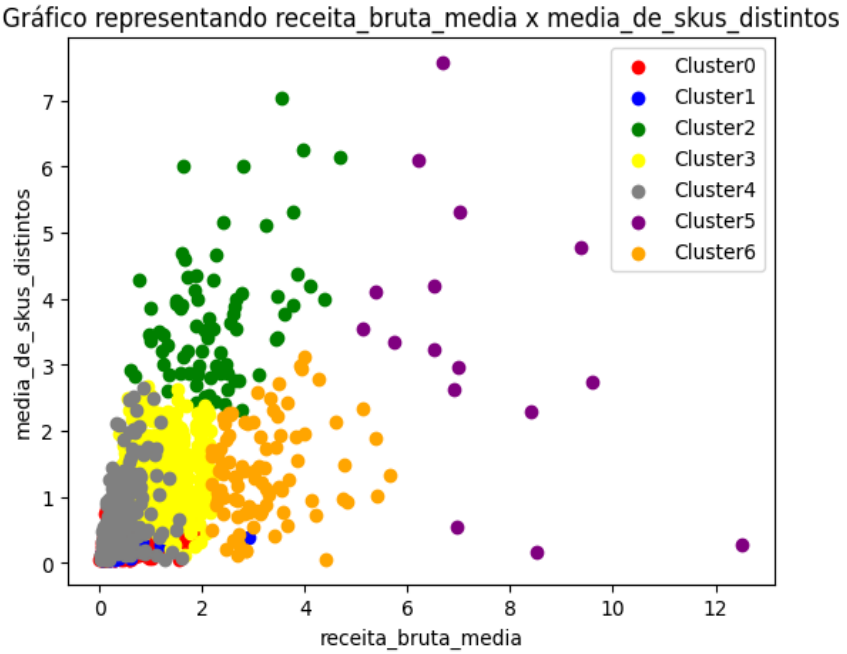
Ao fim deste tópico pode-se notar que, quão melhor concentrados estiverem os clientes ao redor da média de um KPI para seu respectivo cluster, melhor clusterizados estão os clientes, sendo o cenário ideal de separação entre os clusters o cenário em que todos os clientes estão extremamente próximos uns dos outros. Entretanto, aceita-se essa dispersão por não se desejar acrescentar mais clusters pois, como indicado pelo Elbow Method, esta adição, embora permitisse reduzir o erro total da clusterização, isto é, diminuir o desvio padrão dos clientes ao redor das médias dos clusters, não trariam mais uma redução tão significativa para mais de K=7 clusters para que fosse justificado esse acréscimo.

4.2.12 Análise 2D da distribuição dos clientes de cada cluster entre os KPIs

Uma outra análise complementar pode ser realizada ao detalharmos melhor os gráficos do tópico 3.3.2.5 que comparam os clientes nos *KPIs* 2 a 2 porém agora de forma clusterizada. Trazendo três daqueles gráficos que cubram cada relação entre os três *KPIs* do projeto, podemos identificar quais os fatores que melhor separam os clientes dos clusters.

No Gráfico 23 temos a análise da distribuição da receita bruta x média de skus distintos entre os clientes dos clusters. Nesse gráfico podemos observar que os clusters 0, 1 e 4 possuem grande interseção nesses dois *KPIs*, não podendo ser suficientemente diferenciados apenas por eles em alguns casos. Na sequência da diagonal principal podemos ver que o Cluster 3 tem performance melhor que os três primeiros clusters e pior do que os demais clusters seguintes nesses dois *KPIs* se diferenciando pela combinação destes. Os Clusters, 2, 5 e 6 por sua vez se destacou obtendo valores mais expressivos para essas métricas.

Gráfico 23: Distribuição da receita bruta x média de skus distintos entre clientes

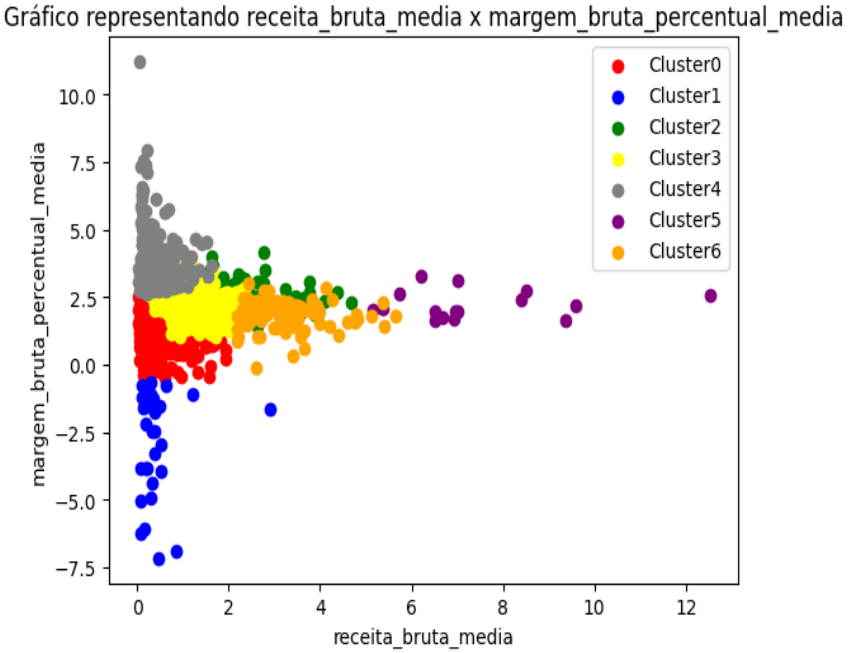


Fonte: Elaboração Própria

De forma adicional ao gráfico anterior, o Gráfico 24 destaca a diferença entre os Clusters 0, 1 e 4 a partir de uma visão da distribuição de clientes com relação à receita bruta x margem bruta que não podiam ser visivelmente diferenciados no gráfico anterior. Neste faz-se

nítida a diferença de performance entre esses clusters de modo que o Cluster 1 destaca-se negativamente por ter performance muito inferior a todos os demais, o Cluster 0 tem performance de margem bruta intermediária e o Cluster 4 possui a melhor margem dentre esses. Os Clusters 2, 3, 5 e 6 por sua vez possuem patamar muito similar de margem bruta, sendo diferenciados nesta visualização principalmente pela receita bruta no eixo x.

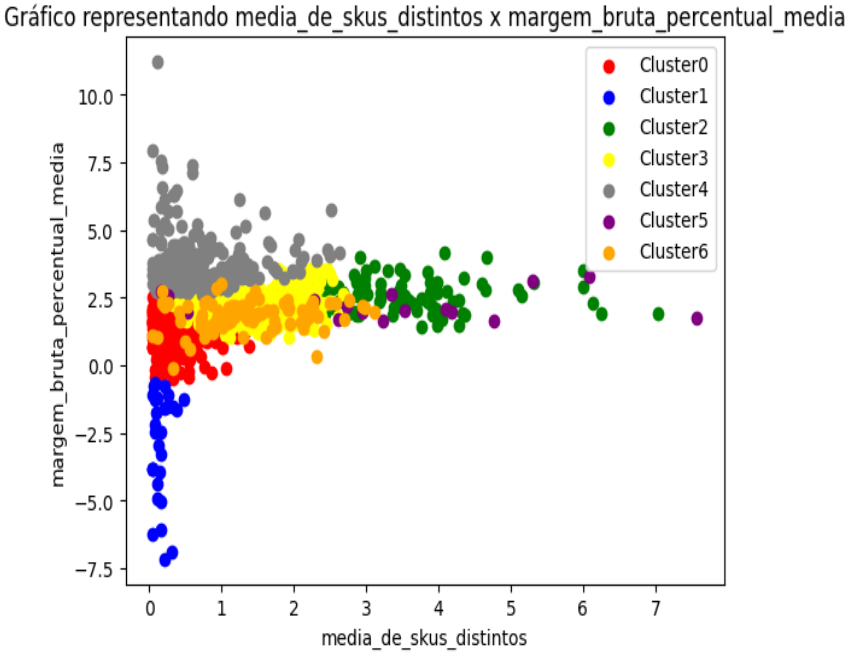
Gráfico 24: Distribuição da receita bruta x margem bruta entre clientes



Fonte: Elaboração Própria

No Gráfico 25 por fim, é trazida a distribuição dos clientes com relação à média de skus distintos x margem bruta adicionando informações complementares aos dois gráficos anteriores.

Gráfico 25: Distribuição da média de skus distintos x margem bruta entre clientes

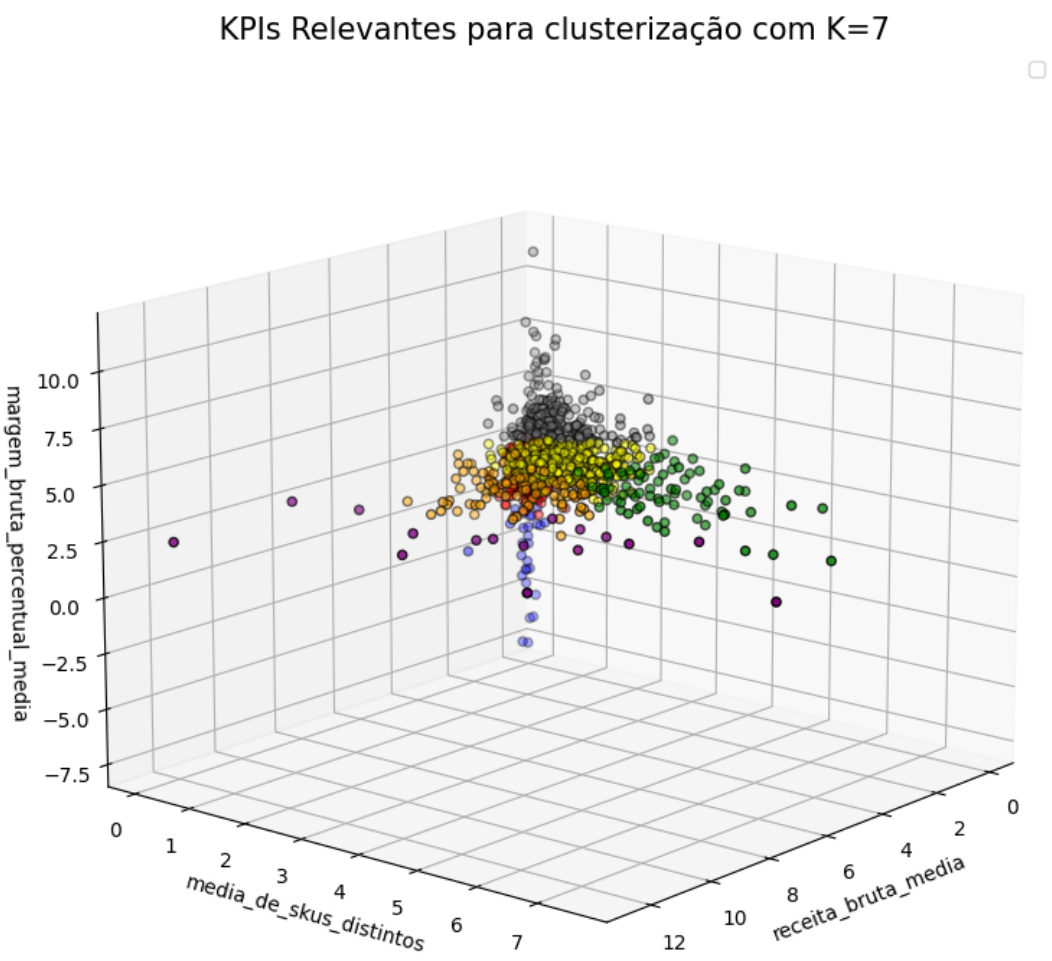


Fonte: Elaboração Própria

4.2.13 Análise 3D da distribuição dos clientes de cada cluster entre os KPIs

Por fim, é trazida uma última visão alternativa na qual é exposta no Gráfico 26 a relação de todos os três clusters. Essa visão traz informações de forma muito mais completa por trazer todos os três KPIs em um único gráfico, porém se configura com uma visão um pouco mais complexa de interpretar por possuir muitas informações neste mesmo gráfico.

Gráfico 26: Distribuição dos KPIs relevantes entre os clusters de clientes



Fonte: Elaboração Própria

Por fim, pode-se concluir que nesta etapa do projeto foram definidos 7 clusters distintos de clientes facilmente identificáveis com base nos KPIs de interesse do projeto. Na próxima etapa estes clusters serão caracterizados de modo que serão conferidas nomenclaturas mais didáticas para a explicação destes clusters aos *stakeholders* do projeto e serão também classificados os clusters em bom, médio ou ruim de acordo com a atratividade destes para a empresa.

4.3 Caracterização dos Clusters de Clientes

Tendo posse das análises individuais da performance de cada cluster em relação a cada KPI é possível determinar uma caracterização clara para os clientes de cada cluster (Ilustração 11). Essa caracterização será dada por meio de nomenclaturas didáticas expostas na segunda coluna da tabela e será a base para as apresentações do projeto.

Ilustração 11: Caracterização dos clusters

Perfil e Jornada dos Clientes
Resumo da Distribuição Geral dos Clientes em Clusters

Cluster	Nome	# Clientes	Receita Bruta	# SKUs	Margem Bruta
0	Oportunista	510	BAIXA	BAIXA	BAIXA
1	Tóxico	29	BAIXA	MUITO BAIXA	MUITO BAIXA
2	Penetrado em Sortimento	79	ALTA	MUITO ALTA	MÉDIA
3	Padrão	422	MÉDIA	MÉDIA	MÉDIA
4	Emergencial	261	BAIXA	BAIXA	ALTA
5	Penetrado em Receita	16	MUITO ALTA	ALTA	MÉDIA
6	Baixa Performance	90	ALTA	MÉDIA	BAIXA

Fonte: Elaboração Própria

O Cluster 0 foi nomeado de Cluster Oportunista, pois concentra clientes que fazem compras muito pontuais, possuindo tanto receita bruta quanto variedade de skus comprados muito reduzida. Este fator faz com que seja notório que ele não abasteça sua loja com a empresa, mas sim realize pequenas compras de oportunidade direcionadas por bons preços promocionais, como pode indicar a baixa margem bruta dos clientes neste cluster.

O Cluster 1 foi nomeado de Cluster Tóxico pois concentra os clientes com pior performance da empresa. Neste grupo estão clientes de alta nocividade para a lucratividade da empresa por possuírem a pior margem bruta dentre todos os clusters. Além disso, o volume de compras desses clientes é baixo o suficiente para, na maioria dos casos, não compensar o

custo de oportunidade de direcionar esforços para atender esses clientes ao invés de atender outros clientes.

O Cluster 2 representa um dos melhores clusters para a empresa no momento, sendo apelidado de Penetrado em Sortimento. Isto porque a empresa consegue vender altos volumes de skus distintos para este cliente, mesmo que a receita bruta mensal dele não seja a mais alta entre todos os grupos. A margem bruta por sua vez é bastante interessante e consegue adicionar grande contribuição para a rentabilidade do modelo de negócios da empresa.

O Cluster 3 representa uma concentração de clientes com performance mediana em todos os *KPIs* analisados e, portanto, foi nomeado de Cluster Padrão. Este grupo tem um destaque para seu volume de clientes o que faz com que, embora este grupo não possua destaque expressivo em nenhum *KPI* ele seja bastante relevante por concentrar, somando todos seus clientes, uma parcela bastante significativa dos negócios da empresa.

O Cluster 4 por sua vez representa um grupo bastante interessante devido a concentrar os clientes com a melhor margem bruta da empresa. Existe um comportamento esporádico de compra desses clientes que faz com que ele realize apenas algumas pequenas compras em receitas e em variedade de skus, embora sejam compras muito rentáveis. Portanto, este grupo foi chamado de Cluster Emergencial, pelo cliente fazer pequenas compras de emergência em que não toma tanta atenção com o preço pago.

O Cluster 5 foi denominado de Penetrado em Receita. Como pode-se inferir pelo nome, este cluster tem os clientes para os quais a empresa consegue vender a maior receita bruta mensal. Destaca-se também a alta variedade de skus comprados para este grupo.

Por fim, o Cluster 6 é o grupo de clientes de Baixa Performance, porque fazem muitas compras, representando uma das receitas brutas mais altas da empresa ao custo de uma das piores margens brutas. Ou seja, os clientes neste cluster são potencialmente muito nocivos para a companhia por comprarem em grandes quantidades itens que ferem a margem da empresa.

Nesta seção podemos também classificar os clusters em bons, médios ou ruins, de acordo com a atratividade do perfil de compra dos clientes contidos nestes clusters para a empresa cliente do projeto, tomando como base a estratégia de negócio da companhia. Dentre os clusters considerados ruins, é possível identificar que o Cluster 1 apresenta baixa performance para os 3 *KPIs*, sendo o pior cluster visualmente identificável. O Cluster 0 possui performance similar ao Cluster 1 com relação à receita bruta, porém destaca-se deste por possuir variedade de skus comprados e margem bruta significativamente piores.

Com relação aos clusters considerados de performance média, podemos identificar os Clusters 3 e 6. O Cluster 3 é o Cluster de clientes Padrões, tendo performance média em todos os três indicadores elencados. E o Cluster 6 possui diferente performance entre os indicadores, tendo receita bruta e variedade de skus respectivamente alta e média, não poderia ser considerado um cluster de performance ruim, porém devido à baixa margem bruta, não atinge o patamar de clusters bons para a empresa.

No que se refere aos clusters de alta performance, temos os Clusters 2, 4 e 5. O Cluster 2 representa os clientes Penetrados em Sortimento destacando-se pela alta variabilidade de skus comprados. O Cluster 4 representa os Clientes Emergenciais, se destacando pela elevada margem bruta média dos clientes deste grupo. E o Cluster 5 é composto pelos Clientes Penetrados em Receita, destacando-se especialmente pela elevada receita bruta destes clientes.

4.4 Previsão de Potencial de Compra da base de clientes

Nesta seção do projeto, buscou-se facilitar aos vendedores, que são uns dos principais *stakeholders* do projeto, a realização de uma melhor gestão da sua carteira de clientes. Para tanto, buscou-se definir o Potencial Máximo de Compra de cada cliente da base da empresa. O Potencial Máximo de Compra foi adotado inicialmente como sendo a máxima performance mensal em termos de receita bruta, variedade de skus comprados e margem bruta que o cliente atingiu em um mês do período analisado.

Para a elaboração deste Painel de Controle de Potencial de Compra Máximo dos clientes foi utilizada a ferramenta de *Business Intelligence* da empresa, o Metabase. Inicialmente foram acessadas as informações a respeito dos clientes e dos clusters aos quais estes foram atribuídos na clusterização anterior (Apêndice B).

Na sequência foi calculado o desempenho de cada cliente no mês corrente a partir da tabela *CONTRIBUTION_MARGIN_1_PER_LINE_ITEM*, expondo os resultados das vendas em termos de receita, variedade de skus e margem bruta percentual para o cliente até o momento (Apêndice B).

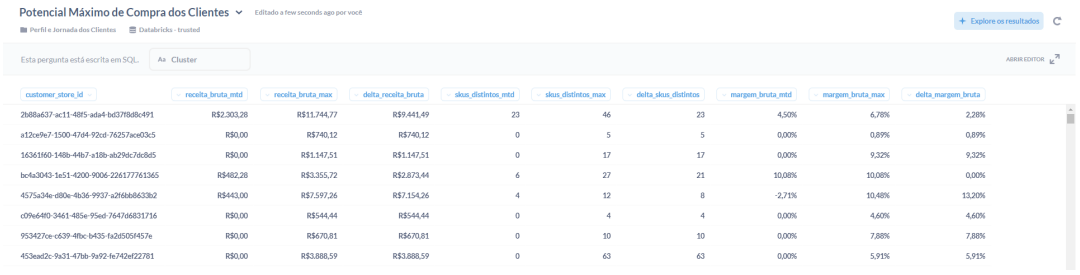
Foi também calculado o Potencial Máximo desses clientes utilizando a mesma tabela mencionada acima (Apêndice B).

Com os resultados dos códigos apresentados no Apêndice B foi possível compilar as informações dos clientes, os clusters desses clientes, a performance atual mensal desses

clientes e o Potencial Máximo de cada um deles. Ainda no Apêndice B pode-se também observar o código que compila todas essas informações de forma unificada.

Essa informação foi disposta aos vendedores por meio de um painel no Metabase demonstrado na Ilustração 12. Para cada cliente, além do Potencial Máximo que ele pode atingir e do resultado atual das vendas para ele no mês corrente, foi apresentado também o delta de performance que é dado pela diferença entre o potencial máximo e o potencial atingido até o momento atual do mês corrente. Este Painel foi construído de modo que fosse possível aos vendedores filtrar os clientes a partir do cluster ao qual eles pertencem.

Ilustração 12: Painel de Potencial Máximo de Compra dos Clientes



Fonte: Elaboração Própria

Dessa forma, os vendedores terão em mãos os clusters dos clientes, a caracterização destes clusters e o Potencial Máximo de Compra dos clientes contidos em cada um destes. Neste estágio faz-se necessário identificar qual seria a melhor distribuição da base de clientes da empresa para auxiliar os vendedores a usarem esta ferramenta da melhor forma possível no intuito de atingir o objetivo do projeto.

4.5 Otimização da Distribuição da Base de Clientes por Perfil de Compra

Nesta etapa do desenvolvimento do projeto será identificada a composição ideal da base de clientes da empresa entre os diferentes clusters de perfil de compra existentes. Sabe-se que os clusters possuem performances distintas com relação aos *KPIs* do projeto. Sabe-se também que alterar a distribuição da base de clientes de modo a aumentar a concentração de clientes de um clusters que performe bem em determinado *KPI* tende a

aumentar a performance geral deste *KPI* para a empresa. Entretanto, é notório que não é possível aumentar ou diminuir indefinidamente os clusters de clientes, haja vista a complexidade de se encontrar e manter apenas clientes dos perfis bons na empresa.

Dessa forma, foi identificado o percentual mínimo e máximo que cada cluster de clientes atingiu no período de apuração analisado (Tabela 8).

Tabela 8: Mínimo e Máximo share de base histórico de cada cluster

Cluster	Mínimo % de Clientes	Máximo % de Clientes
0	18,25%	41,82%
1	0,36%	12,56%
2	5,90%	9,63%
3	17,10%	30,01%
4	13,67%	42,21%
5	0,30%	2,54%
6	2,27%	10,87%

Fonte: Elaboração Própria

Foram definidas como restrições para o otimizador que nenhum cluster poderia ter, na composição otimizada da base, um share maior do que 20% do seu máximo share histórico ou menos que seu mínimo share histórico. Com isso, foram definidas as restrições inferiores e superiores para cada cluster de clientes.

Tabela 9: Restrições para share de base dos clusters no otimizador de base

Cluster	Mínimo % de Clientes	Máximo % de Clientes	Restrição Inferior	Restrição Superior
0	18,25%	41,82%	14,60%	50,18%
1	0,36%	12,56%	0,29%	15,07%
2	5,90%	9,63%	4,72%	11,56%
3	17,10%	30,01%	13,68%	36,01%
4	13,67%	42,21%	10,93%	50,65%
5	0,30%	2,54%	0,24%	3,04%
6	2,27%	10,87%	1,82%	13,05%

Fonte: Elaboração Própria

Para execução do Solver, foram trazidos os valores de receita bruta média, de receita líquida média, de skus distintos médios, de lucro bruto médio e de margem bruta média

construídos anteriormente (Tabela 10). Estes parâmetros servirão para o otimizador identificar quais clientes devem possuir maior representação dentro da base ideal de clientes.

Tabela 10: Potencial Máximo de compra de cada cluster de clientes

Cluster	Receita Bruta Máxima	Receita Líquida Máxima	Skus Distintos Máxima	Lucro Bruto Máxima	Margem Bruta Máxima
0	R$ 2.238,00	R$ 2.116,41	12,27	R$ 124,28	8,68%
1	R$ 1.450,30	R$ 1.403,02	3,62	-R$ 18,53	-2,51%
2	R$ 7.848,93	R$ 7.321,78	84,89	R$ 573,72	10,45%
3	R$ 4.961,87	R$ 4.638,91	36,97	R$ 331,04	9,91%
4	R$ 1.616,77	R$ 1.514,42	18,66	R$ 176,42	15,17%
5	R$ 25.803,24	R$ 24.371,74	78,81	R$ 1.676,99	8,44%
6	R$ 12.602.12	R$ 11.890,75	37,14	R$ 770,24	8,15%

Fonte: Elaboração Própria

A partir do potencial máximo de clientes, foi aplicado um fator de redução de 20%, indicando ser utópico que a empresa conseguisse que todos os clientes performaram em seu melhor potencial simultaneamente, porém factível que conseguisse fazer todos os clientes performaram a 80% de seu potencial máximo simultaneamente. Com base nisso, foi construída uma base de referência para o otimizador que conta com as performances já reduzidas pelo fator de redução aplicado (Tabela 11).

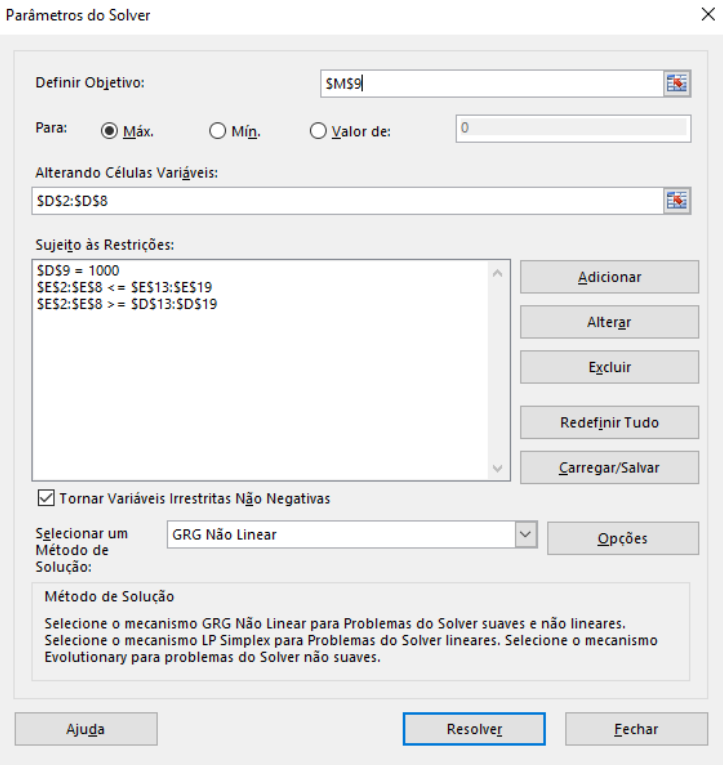
Tabela 11: Potencial Máximo de compra reduzido de cada cluster de clientes

Cluster	Receita Bruta	Receita Líquida	Skus Distintos	Lucro Bruto	Margem Bruta
0	R$ 1.790,40	R$ 1.693,13	9,82	R$ 99,42	6,94%
1	R$ 1.160,24	R$ 1.122,42	2,90	-R$ 14,82	-2,01%
2	R$ 6.279,14	R$ 5.857,42	67,91	R$ 458,98	8,36%
3	R$ 3.969,50	R$ 3.711,13	29,58	R$ 264,83	7,93%
4	R$ 1.293,42	R$ 1.211,54	14,93	R$ 141,14	12,14%
5	R$ 20.642,59	R$ 19.497,39	63,05	R$ 1.341,59	6,75%
6	R$ 10.081,70	R$ 9.512,60	29.71	R$ 616,19	6,52%

Fonte: Elaboração Própria

Por fim, foi aplicada uma restrição de que o número de clientes total da base deveria ser de 1.000 clientes, para tratar de uma base de tamanho hipotético de mais fácil compreensão pelos *stakeholders*, porém com fácil extrapolação de aplicação para a base real da empresa mantendo-se a proporção de clientes em cada cluster. No otimizador foi adotado como parâmetro a ser otimizado o lucro total da base de clientes. O Solver desenvolvido pode ser consultado abaixo na Ilustração 13.

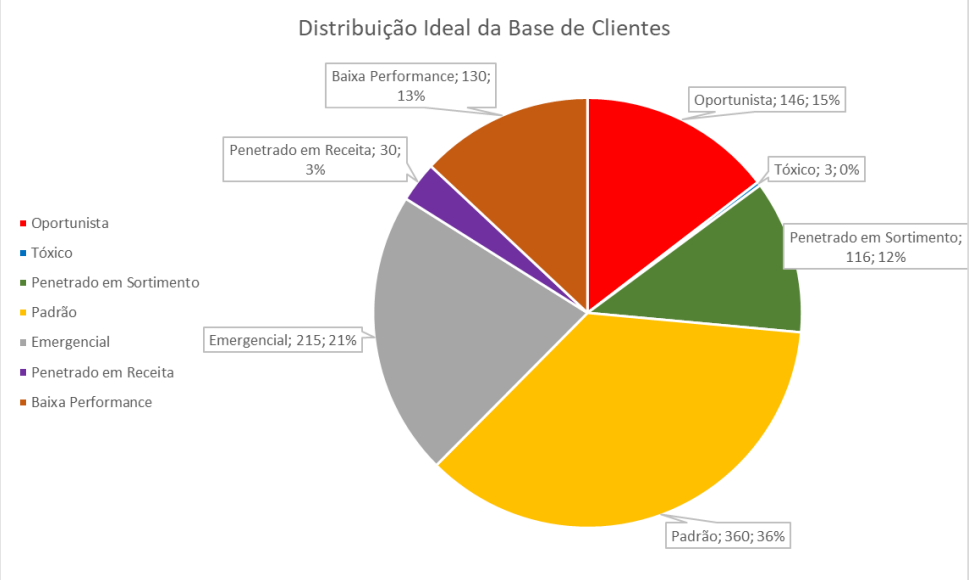
Ilustração 13: *Solver* que otimiza a distribuição da base de clientes da empresa



Fonte: Elaboração Própria

Com a execução do *Solver*, obteve-se o resultado apresentado no Gráfico 27 como sendo a distribuição ideal da base de clientes da empresa entre os clusters identificados. Pode-se observar que os valores das porcentagem de representação de todos os clusters estão devidamente inseridos nos intervalos fornecidos como restrição e a soma total de clientes para a base hipotética é de fato de 1.000 clientes.

Gráfico 27: Distribuição ideal da base de clientes da empresa



Fonte: Elaboração Própria

Sendo assim, a composição ideal da base de clientes da empresa conta com 15% de Clientes Oportunistas, ~0% de Clientes Tóxicos, 12% de Clientes Penetrados em Sortimento, 36% de Clientes Padrões, 21% de Clientes Emergenciais, 3% de Clientes Penetrados em Receita e 13% de Clientes de Baixa Performance. Para a materialização dessa base ideal, é importante o entendimento de quais são os fluxos de transição de clientes entre clusters para a concretização de planos de ação eficientes.

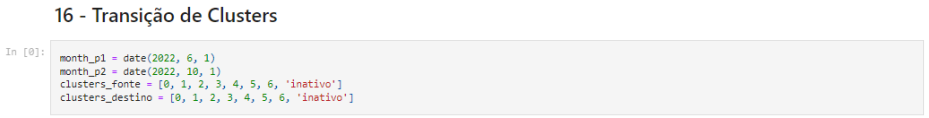
4.6 Análise de transição de clientes entre Clusters

Uma vez identificados os clusters de clientes em que a base da Varejo ABC pode se diferenciar e determinada a composição ideal da base de clientes entre esses clusters, foi identificado como prioridade junto à empresa cliente do projeto a compreensão da transição dos clientes entre os clusters no período analisado. Dessa forma, tendo sido treinado o modelo que permite a obtenção dos resultados apresentados acima, cada cliente foi classificado em um determinado cluster a cada mês de acordo com a performance dele em receita bruta, variedade de skus e margem bruta no respectivo mês, considerando-se também para análise nesta seção o período entre junho de 2022 e outubro de 2022.

Dessa forma, seguiu-se a criação de um método comparativo que permitiu-se analisar a transição do cliente do mês inicial da análise (junho) para o mês final da análise (outubro) considerando os clusters de 0 a 6 e um grupo de clientes inativos, isto é, que não fizeram

pedidos no mês em questão. Os parâmetros utilizados nesta seção podem ser identificados na Ilustração 14.

Ilustração 14: Parâmetros para análise da transição de clientes entre clusters



Fonte: Elaboração Própria

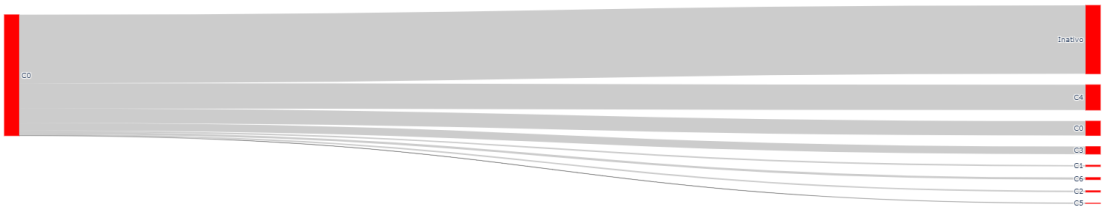
Em posse destes parâmetros, foi construída a função *get_transition_summary* (Apêndice C) que retorna um dataframe contendo o cluster de origem, o cluster de destino e a quantidade de clientes que transacionaram entre estes clusters. Este formato fez-se necessário para a etapa seguinte que envolve a criação de gráficos que permitam a correta visualização deste efeito transitório na base de clientes da empresa.

Com o dataframe obtido através da função get_transition_summary foi possível construir a função *plot_sankey_chart* (Apêndice C) através da qual é possível criar gráficos personalizados para cada cluster de modo que o gráfico mostra do lado esquerdo os clientes que estavam classificados no início da análise (junho de 2022) neste cluster e do lado direito os clusters para os quais esses clientes se destinaram após decorridos os meses da análise.

Na sequência, serão apresentados e discutidos os resultados obtidos através da aplicação da função *plot_sankey_chart* apresentada no Apêndice C para cada um dos clusters definidos.

No Gráfico 28 é possível observar a transição dos clientes do Cluster 0. Desses clientes denominados Clientes Oportunistas, identifica-se que a maioria deles virou inativo no período, indicando a eliminação de clientes deste perfil considerado ruim da base da empresa. Além disso, houve uma parcela de clientes destinados ao Cluster 4 (Clientes Emergenciais) que indica que foi possível aumentar nesta parcela de clientes de maneira significativa a margem bruta das compras realizadas. Ademais, não houveram grandes surpresas em relação aos demais clusters de destino deste grupo de clientes no período.

Gráfico 28: Transição dos clientes do Cluster 0 de Junho a Outubro



Fonte: Elaboração Própria

No Gráfico 29 vemos os clientes do Cluster 1 transicionando no período. A primeira informação relevante é de que os Clientes Tóxicos deste grupo apenas transacionaram a uma gama muito seleta de clusters. Sendo a maior parte deles inativado, como esperado para os clientes de clusters de baixa performance, houveram parcelas de clientes transicionando também para os Clusters 0 e 4, indicando que foi possível melhorar a margem bruta de alguns clientes deste cluster, sem no entanto gerar melhorias significativas na receita bruta e na variedade de skus compradas por estes. Este comportamento indica certa resistência na transição dos clientes deste cluster para níveis mais representativos no share de compras da empresa.

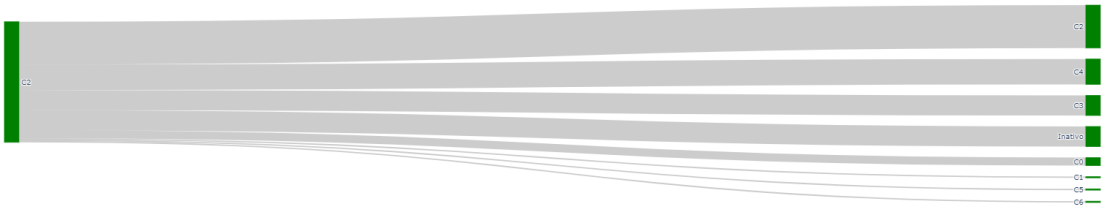
Gráfico 29: Transição dos clientes do Cluster 1 de Junho a Outubro

Fonte: Elaboração Própria

A transição dos clientes do Cluster 2 pode ser observada no Gráfico 30. Ressalta-se que por este cluster compreender uma classificação considerada boa devido a centralizar clientes penetrados em sortimento era desejável que este cluster mantivesse clientes em

grupos de boa performance, idealmente melhorando os parâmetros de receita bruta e de margem bruta que não são a grande força deste cluster, sem deixar que haja uma perda significativa na sua grande força, que é a variedade de skus. Observa-se que, de fato, a maioria dos clientes permanece no próprio Cluster 2 ao fim da análise, com alguns clientes transicionando de forma significativa para o Cluster 4, que indica melhoria na margem bruta e redução na variedade de skus comprados simultaneamente.

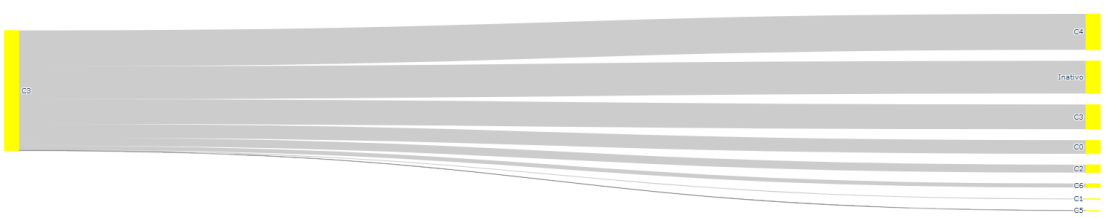
Gráfico 30: Transição dos clientes do Cluster 2 de Junho a Outubro



Fonte: Elaboração Própria

O Cluster 3 concentra os clientes chamados de Clientes Padrão e sua transição pode ser observada no Gráfico 31. Ressalta-se que devido ao número elevado de clientes contido neste cluster, é bastante significativa a transição deste de modo que um movimento de tendência de melhoria da classificação destes cluster indica um movimento de melhoria da base geral da empresa e vice-versa. Os três destaques de transição vão para as migrações para os grupos de clientes para os Clusters 4 e Inativos e para a grande quantidade de clientes que se mantiveram no Cluster 3. A inativação de clientes demonstra certa preocupação pois são clientes bastante relevantes para a empresa.

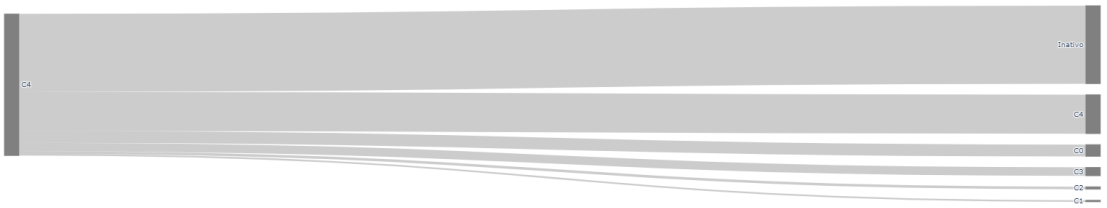
Gráfico 31: Transição dos clientes do Cluster 3 de Junho a Outubro



Fonte: Elaboração Própria

No Gráfico 32 destaca-se um dos cenários de transição mais preocupantes, que é a inativação de uma grande quantidade de clientes do Cluster 4 que, embora, consuma baixa receita bruta e baixa variedade de skus, representa uma força muito significativa para a empresa devido a alta rentabilidade dos clientes deste grupo. Ainda, parte significativa deste cluster se manteve no próprio Cluster 4.

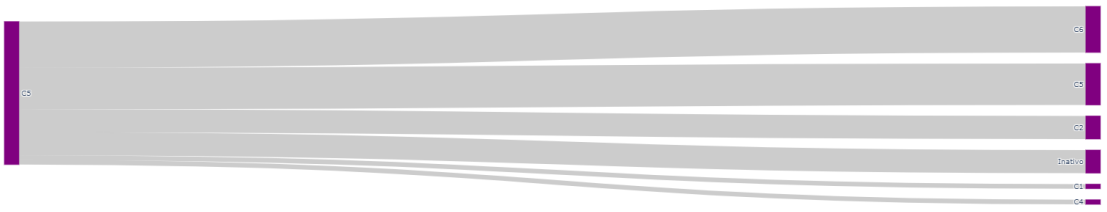
Gráfico 32: Transição dos clientes do Cluster 4 de Junho a Outubro



Fonte: Elaboração Própria

O Cluster 5 teve clientes com redução na sua receita bruta e na sua margem bruta, indicados pela transição de uma porção significativa de clientes para o Cluster 6. Além disso, muitos clientes foram mantidos no próprio Cluster 5 ou migraram para o Cluster 2 (Gráfico 33), que são ambos grupos de clientes muito desejáveis para a empresa.

Gráfico 33: Transição dos clientes do Cluster 5 de Junho a Outubro

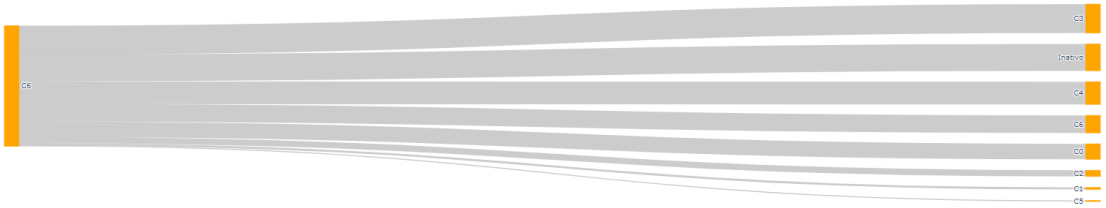


Fonte: Elaboração Própria

O último cluster de clientes propriamente dito é o Cluster 6, que é composto pelos clientes denominados de Baixa Performance por consumiram muito com baixa margem bruta.

Neste grupo existem três destaques transitórios negativos, que é a transição de clientes deste cluster para o grupo de clientes inativos, que indica que foram perdidos alguns clientes prejudiciais para a empresa, e a transição de clientes para os Clusters 3 e 4, que indicam melhoria na performance deste grupo de clientes (Gráfico 34).

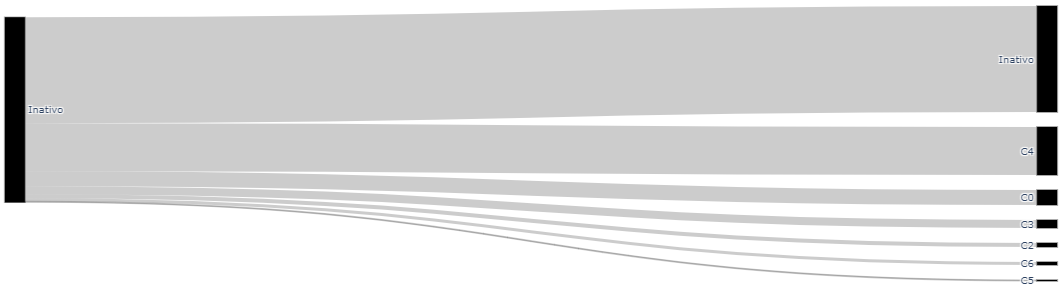
Gráfico 34: Transição dos clientes do Cluster 6 de Junho a Outubro



Fonte: Elaboração Própria

Por fim, observa-se no Gráfico 35 que muitos dos clientes do espaço amostral analisado que começaram o período inativos de fato acabaram o período inativo. Este fator indica que não houve sucesso no período analisado no que se refere à rampagem de novos clientes adquiridos em volume significativo para clusters de alta performance.

Gráfico 35: Transição dos clientes Inativos de Junho a Outubro



Fonte: Elaboração Própria

Nesta seção, pôde-se compreender melhor sobre a transição dos clientes entre clusters, de modo que foram destinados planos de ações aos times operacionais para compreender quais foram os fatores mais significativos para cada uma das transições mais importantes

entre clusters. Com este entendimento, será possível à empresa possuir maior controle sobre essas transições, de modo a incentivar transições benéficas para a sustentabilidade do modelo de negócios e de evitar transições maléficas para este. Resta ainda identificar quais os atributos de um cliente são mais significativos para a classificação deste em cada perfil de compra, facilitando o entendimento da classificação dos clientes nos clusters e dos movimentos transitórios entre os clusters.

4.7 Identificação dos atributos de maior correlação com o Perfil de Compra

Após definidos os clusters em que os clientes da Varejo ABC se dividem e analisar as transições destes clientes entre os clusters, foi essencial a análise dos atributos que possuem maior impacto na classificação dos clientes em cada clusters. Desse modo, serão levantados em conta tanto atributos não intrínsecos ao cliente que tangem seu comportamento de compra em seu histórico de relacionamento com a empresa - que serão aqui chamados de atributos não determinísticos - quanto atributos intrínsecos aos clientes, como seu perfil de loja e sua localização geográfica - que serão aqui chamados de atributos determinísticos.

Será utilizada a Matriz de Correlação de Pearson para identificar a correlação entre cada um destes atributos com o cluster do cliente e com os três KPIs que foram utilizados na clusterização. O intuito desta etapa é identificar os atributos com maior correlação, seja ela positiva ou negativa, com cada um dos clusters de clientes e com os KPIs alvo do projeto.

3.2.1.1 Atributos Não Determinísticos

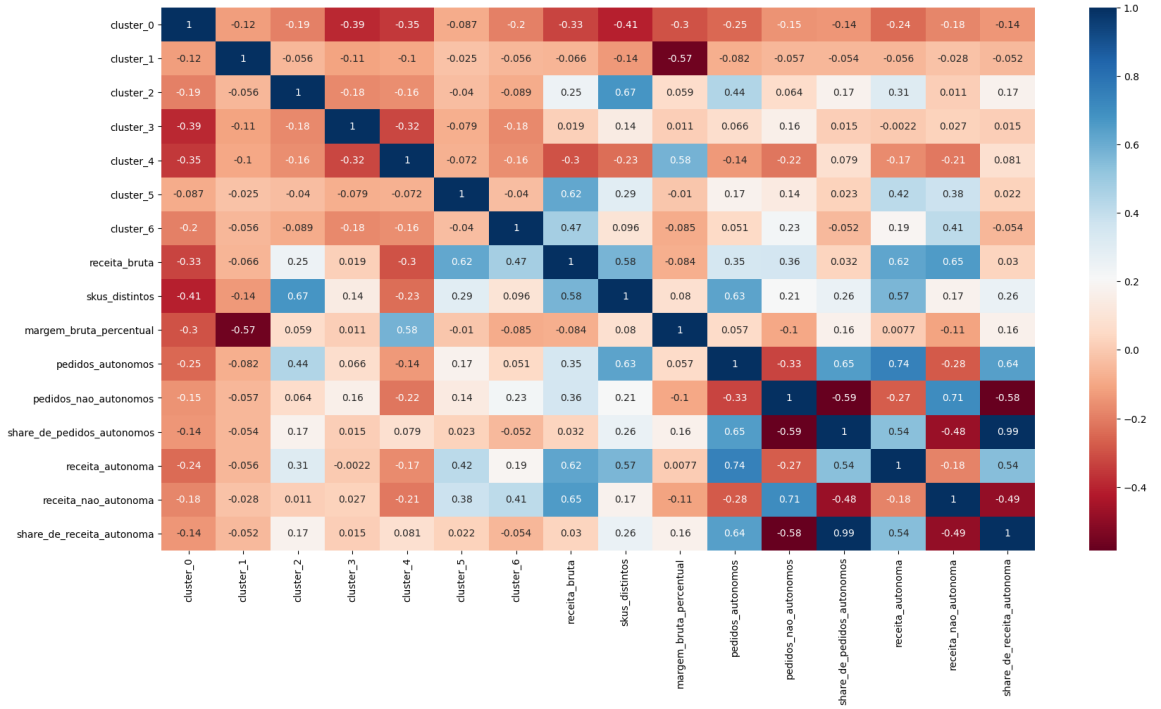
Para a melhor visualização da correlação de cada um dos atributos, as análises desta seção foram divididas em seis matrizes de correlação diferentes. A primeira delas concentra atributos relacionados à autonomia de compra do cliente - ou seja, o quanto ele compra sem a necessidade de auxílio de um vendedor através do site. Na segunda matriz serão analisadas a correlação entre o share de compras a preço promocional e a preço regular que o cliente realiza. A terceira matriz traz uma análise da correlação entre a receita de cada categoria de produtos e as variáveis desejadas. A quarta matriz tem uma visão bastante similar à terceira matriz, mas traz a participação de cada categoria em termos de share do total de pedidos e de receita ao invés de valor bruto. A quinta matriz por sua vez traz atributos que dizem respeito ao perfil de pedido realizado pelo cliente, como a frequência de pedidos e o ticket médio destes. A sexta matriz por sua vez analisa o consumo de crédito fornecido aos clientes de cada

grupo. Por fim, a sétima matriz analisa a correlação do histórico de inadimplência do cliente com sua classificação em clusters.

No Gráfico 36 é possível identificar cinco correlações mais significativas positivas. Nominalmente, se correlacionam positivamente o número de pedidos autônomos e a classificação do cliente no cluster 2; a receita autônoma e a classificação do cliente no cluster 5; a receita não autônoma e a classificação do cliente no cluster 6; a receita não autônoma e a classificação do cliente no Cluster 5; a receita não autônoma e a classificação dos clientes no cluster 6; e a receita autônoma e a classificação do cliente no Cluster 2. Com esta primeira análise primordial, as principais conclusões são de que os clientes do Cluster 2 possuem tanto mais pedidos quanto mais receita autônoma, os clientes do Cluster 5, penetrado em receita, tendem a ser clientes que compram elevada receita tanto de forma autônoma quanto não autônoma e o Cluster 6 tende a concentrar um perfil de compra menos autônomo. Além disso, a receita autônoma tem grande correlação com a variedade de skus que o cliente compra da empresa.

Ainda, podem ser identificadas suas correlações negativas de certa representatividade. A classificação de um cliente no Cluster 0 se correlaciona negativamente tanto com o número de pedidos quanto com a receita autônoma. E os clientes pertencentes ao Cluster 4 se correlacionam negativamente tanto com a quantidade de pedidos quanto com a receita não autônoma.

Gráfico 36: Correlação entre variáveis de autonomia e clusterização

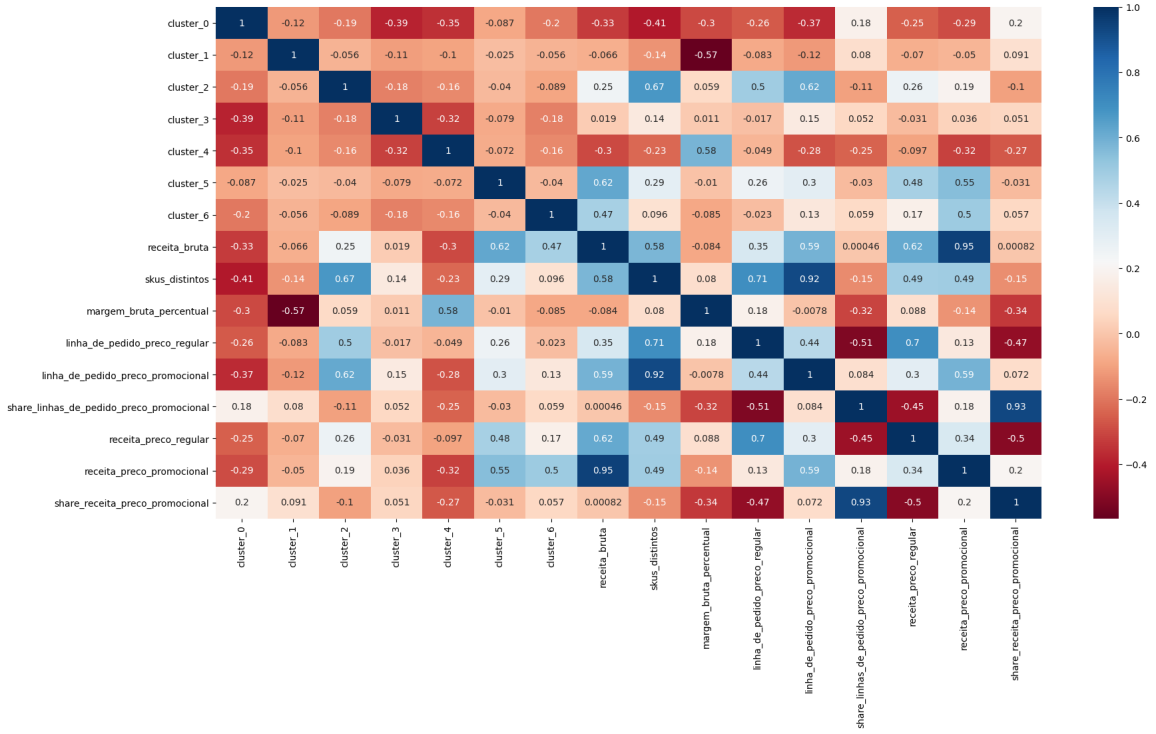


Fonte: Elaboração Própria

No Gráfico 37 podemos observar quatro correlações positivas significativas entre o share de compras a preço promocional e a preço regular de um cliente e a sua atribuição nas variáveis de interesse. O Cluster 5 possui forte correlação positiva tanto com a receita a preços promocionais quanto com a receita a preços regulares, indicando que este cluster de elevada receita não apresenta distinção clara com relação ao modelo de compra predominante, comprando em ambos os casos da empresa. O Cluster 6 possui maior correlação com a receita vinda de promoções, indicando-se um cluster com maior foco em ofertas do que no abastecimento geral de sua loja.

Já o Cluster 2 apresenta também uma correlação positiva entre as quantidades de linha de pedidos (cada produto comprado em um pedido) tanto promocionais e regulares com sua classificação neste cluster, com certa predominância para as linhas de pedido promocionais, indicando que sua penetração em sortimento ocorre nos dois modelos de precificação mas é predominante entre as promoções. Em termos de correlações negativas, os Clusters 0 e 4 são inversamente correlatos à receita advinda de promoções, o que pode indicar perfis menos focados em compras a preços agressivos nestes grupos.

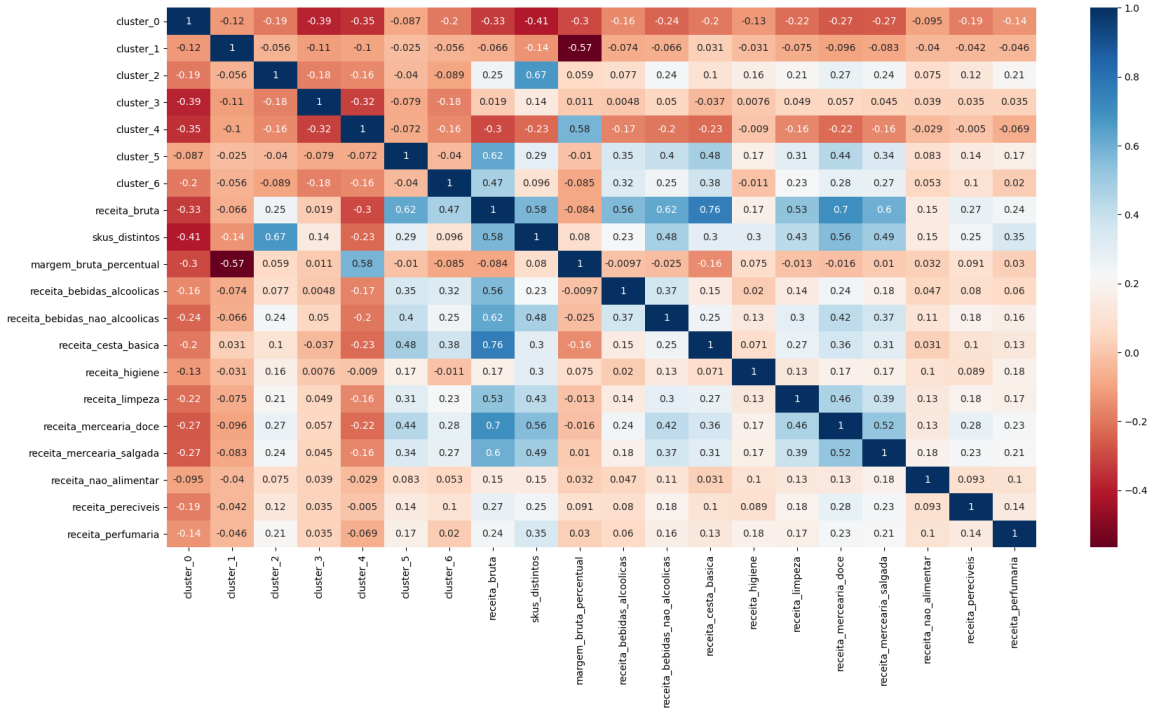
Gráfico 37: Correlação entre variáveis de share de promoções e clusterização



Fonte: Elaboração Própria

Em termos dos atributos focados na distinção da receita da empresa entre as categorias de produtos no Gráfico 38, identifica-se que o Cluster 5 possui seu destaque em receita impulsionado principalmente pelas categorias de cesta básica, mercearia doce e bebidas não alcoólicas. O Cluster 6 se correlaciona positivamente com as categorias de cesta básica, bebidas alcoólicas, mercearia doce e mercearia salgada. O Cluster 4, por sua vez, que apresenta elevada margem bruta, possui uma correlação inversamente proporcional com as categorias de cesta básica e de mercearia doce. Um último destaque se dá à correlação entre cesta básica e mercearia doce com receita bruta, indicando que os clientes que compram maiores volumes dessas categorias tendem a comprar um maior volume de receita no geral.

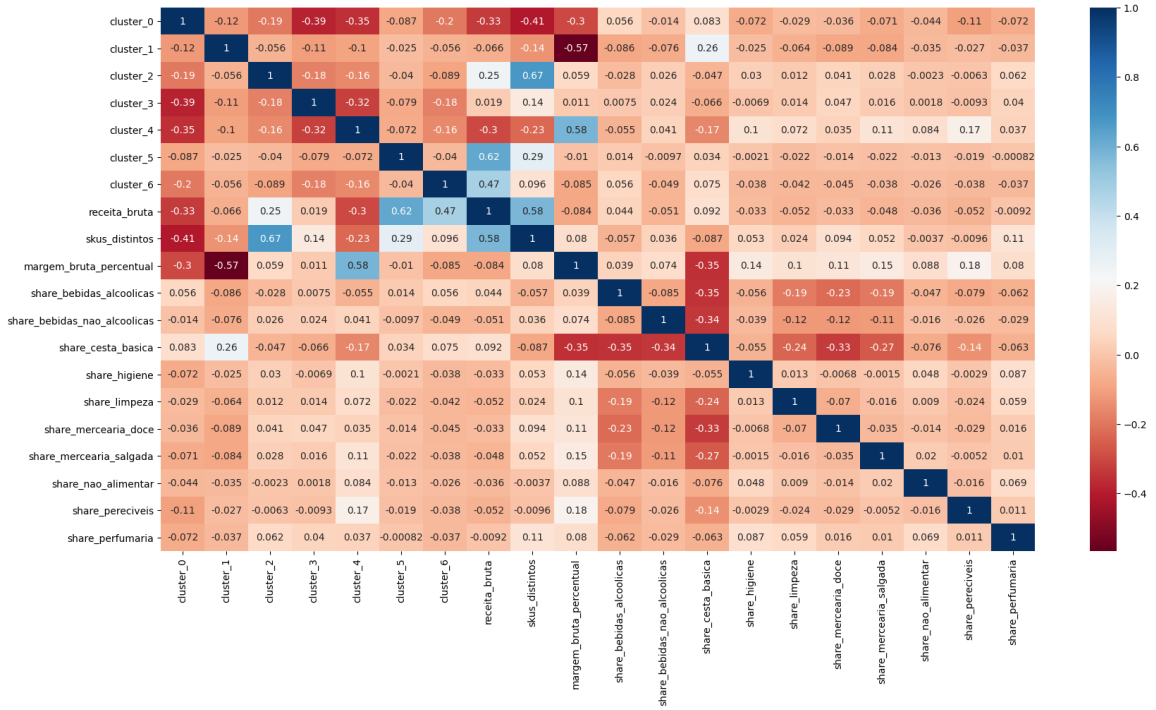
Gráfico 38: Correlação entre variáveis de receita por categoria e clusterização



Fonte: Elaboração Própria

De maneira complementar ao Gráfico 38, o Gráfico 39 traz a visão da distribuição das vendas nas categorias como share de receita ao invés da receita bruta de cada categoria. Nesta visão, o destaque vai para o Cluster 1 que possui alta correlação positiva com o share de cesta básica deste grupo e para o Cluster 4 que possui alta correlação negativa com esta mesma categoria. Além disso, a margem bruta percentual se correlaciona negativamente com o share de cesta básica nas compras dos clientes, indicando baixa rentabilidade desta categoria.

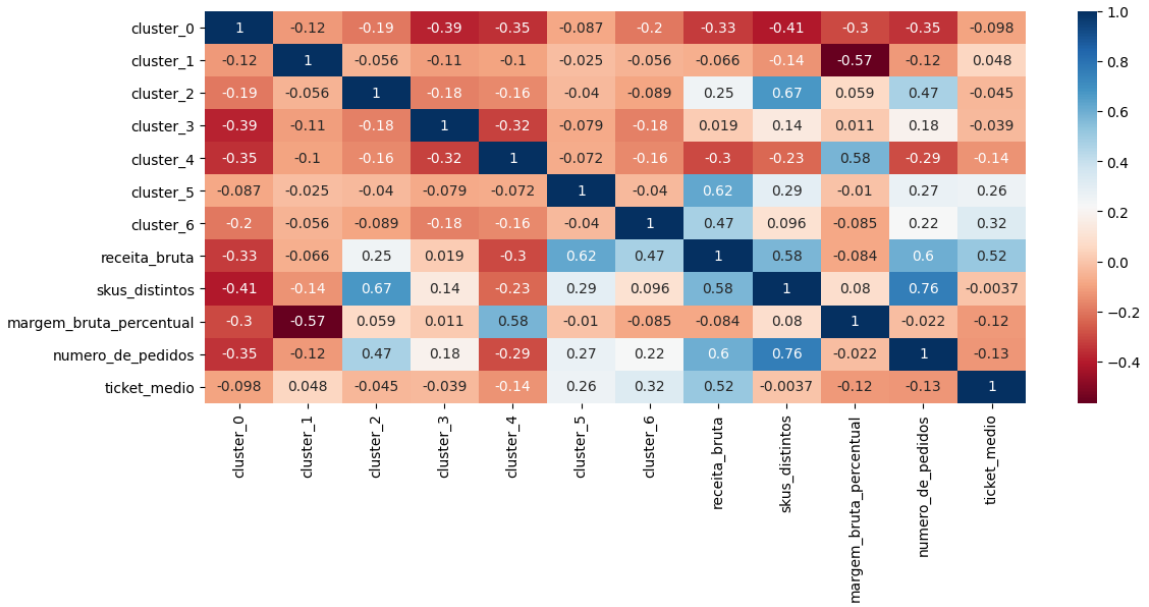
Gráfico 39: Correlação entre variáveis de share de categoria e clusterização



Fonte: Elaboração Própria

No que se refere a perfil de pedidos, foram mapeados os atributos de quantos pedidos o cliente faz no mês e de qual o ticket médio destes pedidos. No Gráfico 40 pode-se observar que o Cluster 2 é o cluster com maior correlação positiva com o número de pedidos mensais realizados, seguido pelos clusters 5 e 6, respectivamente. Em termos de ticket médio dos pedidos, os Clusters 6 e 5 também apresentam grande correlação positiva com este indicador, indicando que os clientes nestes clusters tendem a fazer pedidos maiores. Os Clusters 0 e 4 são os mais inversamente correlacionados com o número de pedidos realizado e não há nesta análise correlações negativas significativas para o ticket médio.

Gráfico 40: Correlação entre variáveis de perfil de pedidos e clusterização

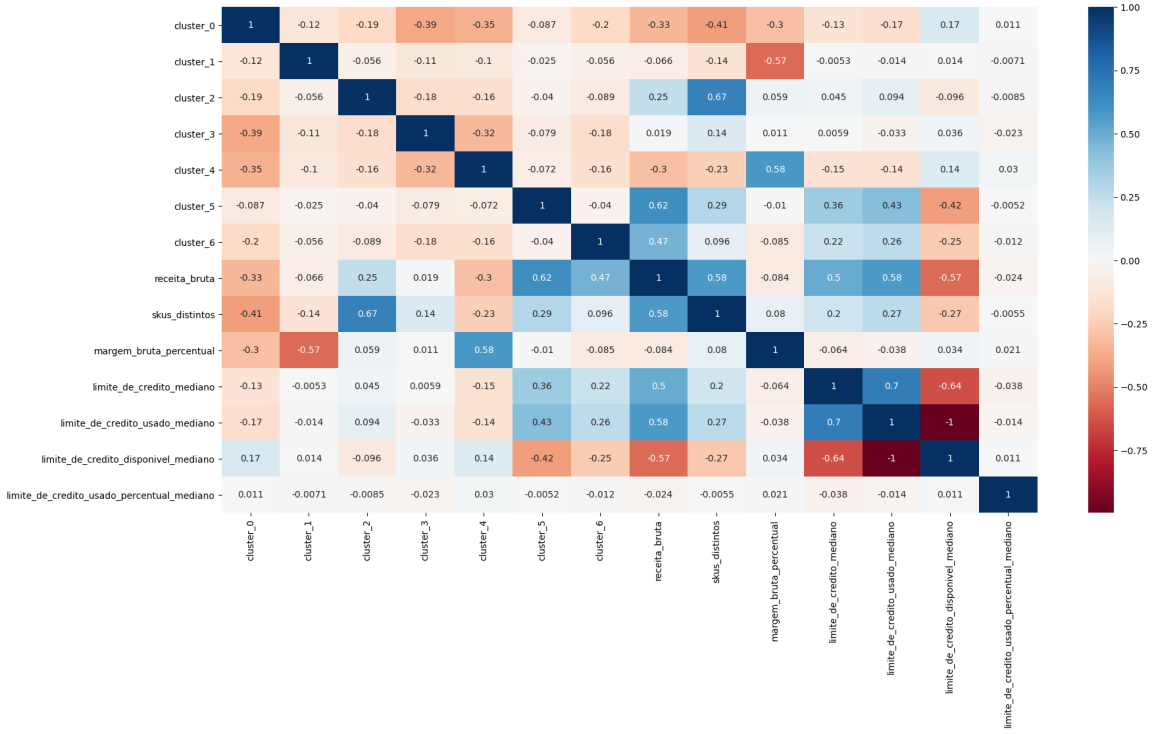


Fonte: Elaboração Própria

O Gráfico 41 analisa a correlação entre o consumo de crédito dos clientes e sua respectiva clusterização. Para esta análise, foram observados os indicadores de limite de crédito total, limite de crédito disponível, limite de crédito utilizado e limite de crédito percentual usado. Para cada um desses três indicadores foram analisadas as performances medianas com respeito a eles.

Os Clusters 5 e 6 se destacam como os clusters com maior correlação com o limite de crédito total e com o limite de crédito utilizado, características aparentemente determinantes para a elevada receita obtida por ambos os grupos. Além disso, o Cluster 5 tem correlação bastante negativa com o limite de crédito disponível, indicando que em geral sobra pouco crédito a ser consumido no mês para este grupo de clientes. A receita bruta também é positivamente correlacionada com o limite de crédito total e com o limite de crédito usado e inversamente correlacionada com o limite de crédito disponível, indicando que clientes com maior limite de crédito e com maior consumo desse limite tendem a ter maiores receitas mensais.

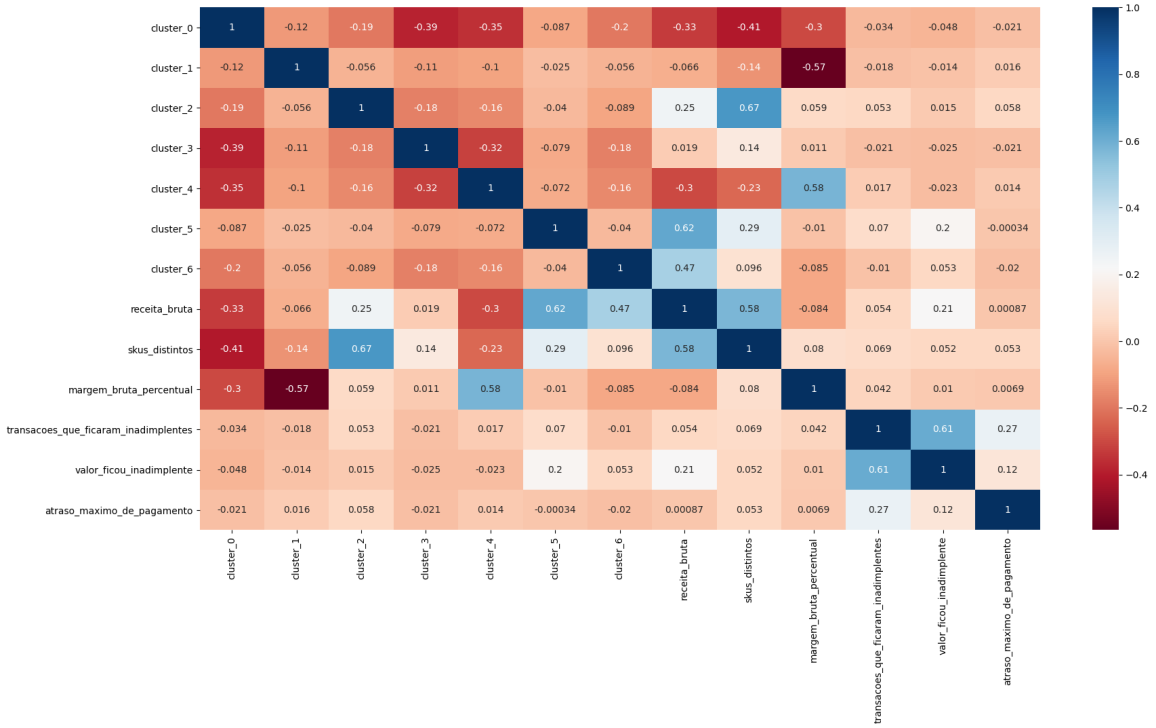
Gráfico 41: Correlação entre variáveis de consumo de crédito e clusterização



Fonte: Elaboração Própria

No Gráfico 42 podem ser analisados os atributos relacionados com a inadimplência dos clientes e sua correlação com o cluster destes. Nesta análise foram considerados o número de transações do cliente que ficaram inadimplentes, o valor total do cliente que ficou inadimplente e o atraso máximo de pagamento de uma transação pelo cliente. Nesta análise, a correlação mais expressiva é a do Cluster 5 com o valor inadimplente, indicando que este perfil tem maior tendência à inadimplência. Não existem outras correlações significativas nesta análise.

Gráfico 42: Correlação entre variáveis de inadimplência e clusterização



Fonte: Elaboração Própria

Com as análises realizadas nesta seção, é possível se obter uma melhor compreensão acerca do que são fatores determinantes para a classificação de um cliente em cada cluster. Nota-se que os clusters com características de destaque mais expressivas, como é o caso dos Clusters 1, 2, 4, 5 e 6 possuem maior facilidade em se identificar fatores correlatos a eles para sua classificação e clusters com performance menos expressivas como os Clusters 0 e 3 possuem menos fatores de forte correlação.

Com este resultado, o time de vendas poderá estudar melhor o perfil dos clientes de cada cluster com três principais finalidades. Primeiramente, será possível compreender melhor a transição dos clientes a partir do conhecimento de como a performance nos atributos mais correlacionados variou no período. Também será possível definir perfis foco para prospecção de clientes em momentos de expansão de base de clientes pela empresa. De modo inverso, será possível identificar perfis de comportamento não desejáveis na empresa que deverão ser desestimulados através da adequação da base de clientes para remover clientes que possuam estas características. Por fim, será possível realizar a gestão da base de clientes ativos de modo a utilizar das alavancas de atributos mais correlacionados com os clusters para melhorar o potencial de compra dos clientes.

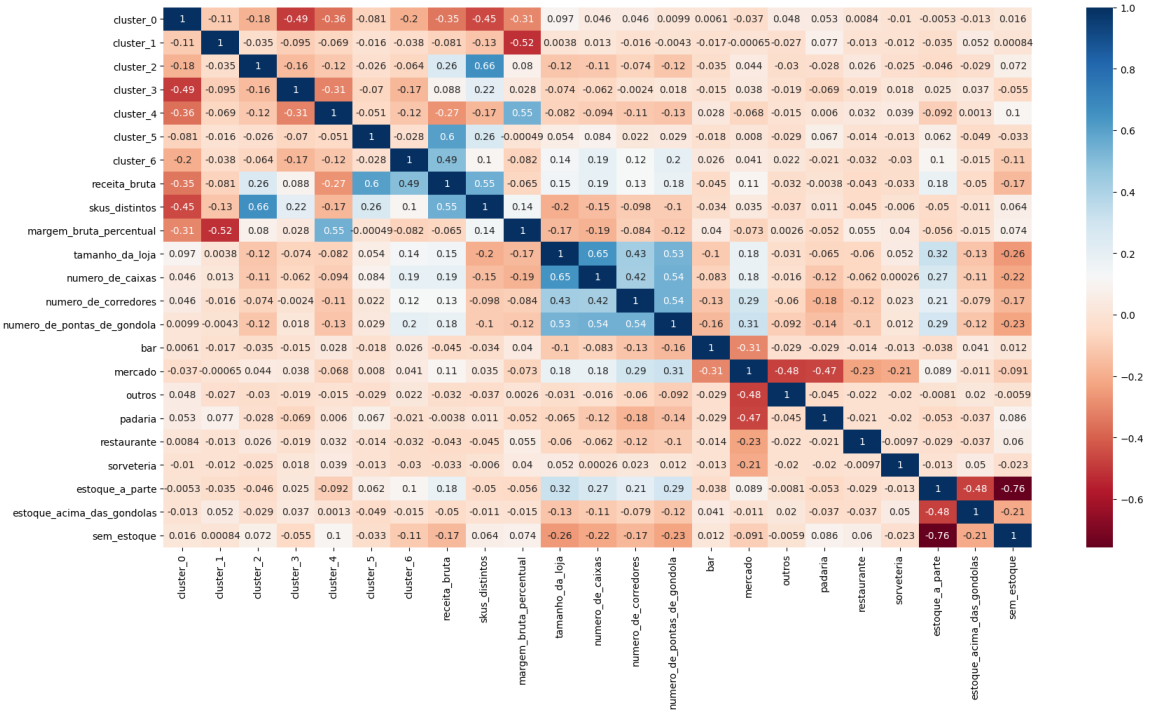
3.2.1.2 Atributos Determinísticos

Os atributos determinísticos foram divididos em variáveis referente ao perfil de loja do cliente e variáveis referentes às características geográficas da região onde o cliente se encontra. Cada grupo de variáveis foi atribuído a uma Matriz de Correlação Própria.

Com relação às variáveis de perfil de loja, foi analisado o segmento da loja, o tamanho da loja, o número de corredores, o número de caixas registradoras, o número de pontas de gôndola existentes e o formato de estoque. As variáveis categóricas (segmento de loja e formato de estoque) foram separadas em diferentes variáveis binárias de modo que cada variável represente uma opção dentre as selecionadas para a variável original, marcando 1 para o pertencimento do cliente a esta variável e 0 para o não pertencimento dele.

Nesta análise, todas as correlações observadas foram fracas, não havendo conclusões precisas de que o perfil de loja possa impactar na clusterização. A correlação mais forte positiva foi dada entre a detenção de um estoque a parte e a receita bruta do cliente, indicando que clientes que têm maior capacidade de armazenamento de seus produtos compram mais da empresa.

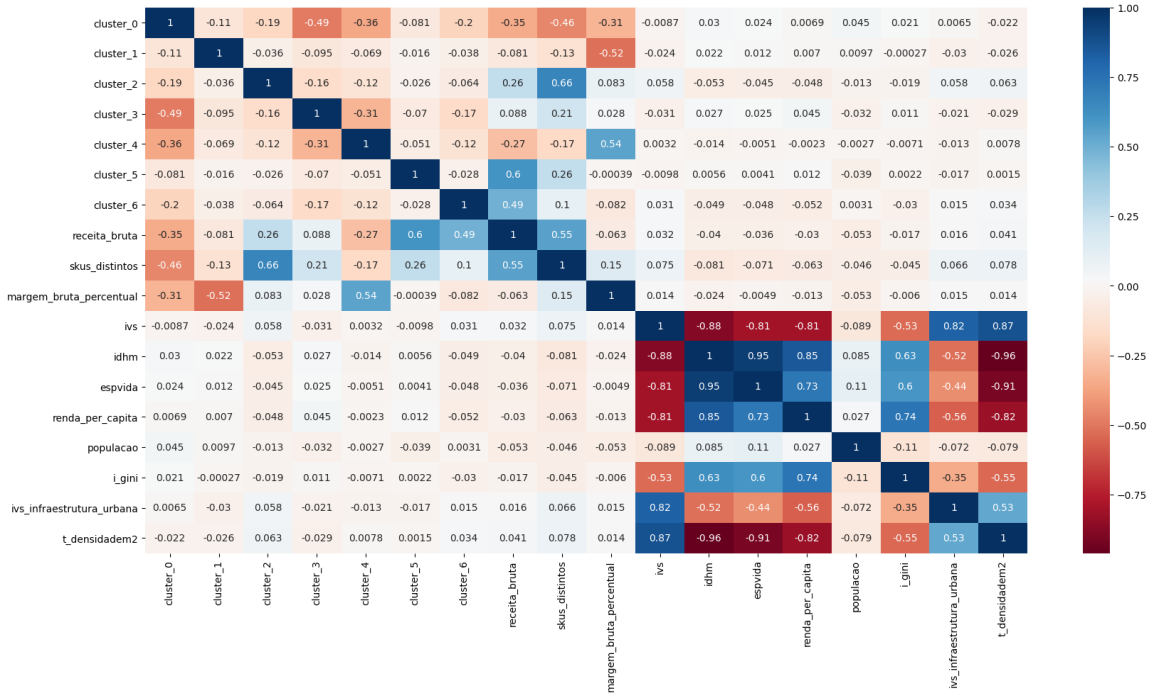
Gráfico 43: Correlação entre variáveis de perfil de loja e clusterização



Fonte: Elaboração Própria

Já no Gráfico 44 foram analisados os aspectos referentes às características geográficas do cliente. Assim como no gráfico acima, para este caso também não foram encontradas correlações significantes entre os aspectos aqui descritos e a clusterização desenvolvida.

Gráfico 44: Correlação entre variáveis de localização geográfica e clusterização



Fonte: Elaboração Própria

Nesta última análise pode-se concluir que os atributos determinísticos do cliente tem pouca relação com o cluster ao qual o cliente foi atribuído.

4.8 Definição de Planos de Ação

A etapa final de desenvolvimento do projeto consiste na consolidação de planos de ação a serem executados futuramente pela empresa cliente. Os planos de ação desenvolvidos tomarão por base todo o detalhamento do projeto realizado na seção 4.1 e a clusterização de clientes elaborada e caracterizada respectivamente nas seções 4.2 e 4.3.

As iniciativas desenvolvidas pelos planos de ação serão planejadas no intuito de se alcançar a distribuição ideal da base ideal de clientes da empresa que foi determinada na seção 4.5 fazendo uso do potencial máximo dos clientes calculado na seção 4.4 e fará uso dos

aprendizados obtidos com a análise da transição de clientes entre clusters da seção 4.6 e da identificação dos atributos correlacionados com a clusterização de clientes desenvolvida na seção 4.7. Desse modo, os planos de ação farão uso de todas as etapas sequenciais desenvolvidas do trabalho para consolidar o objetivo do projeto primordial do projeto e assegurar a alteração da distribuição de perfis de cliente da base atual da Varejo ABC para uma composição que otimize a adesão desta à proposta de valor da empresa.

Após sucessivas reuniões de apresentação, discussão e iteração das diversas entregas parciais deste projeto com os *stakeholders* envolvidos, foram traçados cinco planos de ação pelo autor que serão recomendados para a empresa:

1. Iniciar todas as reuniões matinais diárias do time de vendas acompanhando o delta de realizado mensal vs potencial máximo da carteira de clientes de cada vendedor. Com esta medida, os clientes de maior potencial inexplorado devem ser priorizados diariamente e focados ao longo do dia para que o vendedor otimize seu tempo focando nos clientes de maior potencial restante, trazendo maiores resultados com menores esforços.
2. Construir personas que considerem as matrizes de correlação elaboradas para uma prospecção mais assertiva de clientes pela empresa. Com essa melhor seleção de personas, será possível reduzir o CAC (Custo de Aquisição de Clientes) e aumentar o LTV (Life Time Value) destes tendo em vista que o time de vendas estará mais direcionado a clientes com maior adesão à proposta de valor oferecida pela empresa.
3. Remover da base de clientes da empresa aqueles que estão em clusters classificados como ruins para o modelo de negócios de abastecimento e não foram transferidos com sucesso para clusters de melhor performance no primeiro trimestre do ano que vem. Desta forma será possível evitar a perda do custo de oportunidade de atender clientes não rentáveis, redirecionando recursos para atender clientes mais rentáveis.
4. Incluir no planejamento estratégico do próximo trimestre iniciativas voltadas para reduzir os clusters que precisam ser reduzidos e ampliar os clusters que precisam ser ampliados de acordo com a mudança de distribuição da base de clientes definida. Neste planejamento estratégico deverá ser levado em conta as transições de clusters mais fáceis de serem realizadas de acordo com as análises de transições entre clusters passadas que levem clientes de clusters que precisam ser reduzidos a clusters que precisam ser ampliados.

5. Priorizar no backlog do time de desenvolvimento de software iniciativas voltadas a impulsionar na plataforma da empresa os comportamentos relacionados aos atributos não determinísticos mais relevantes determinados pelas matrizes de correlação. Busca-se assim desenvolver de maneira mais escalável comportamentos benéficos para a rentabilidade da base de clientes.
6. Execução de rotas de pesquisa para visitar clientes de clusters específicos a fim de buscar compreender de forma também qualitativa quais os fatores mais significativos para o pertencimento de cada cliente a cada cluster. Neste plano de ação objetiva-se expandir o potencial das iniciativas de clusterização, caracterização e análise de atributos correlacionados com a clusterização para aspectos mais palpáveis para os vendedores e agregando informações que a empresa ainda não dispõe de dados para possuir.

Com estes planos de ação, é esperado que a empresa atinja em 2023 um patamar muito mais rentável para suas operações. Além disso, é esperado que sejam originadas organicamente pelos times da empresa outros planos de ação proveniente da interpretação dos *stakeholders* dos materiais contidos neste trabalho. Por fim, compete dizer que os planos de ação elencados deverão ser priorizados conforme a disponibilidade de recursos da organização e devem ser iterados conforme os aprendizados obtidos durante sua execução.

5 CONCLUSÕES E CONTRIBUIÇÕES

Após o desenvolvimento do projeto, podem ser compiladas algumas conclusões finais acerca do material desenvolvido. Nesta última seção, serão realizadas conclusões gerais do projeto para uma análise final do atingimento dos objetivos esperados pelo autor e pela empresa. Também serão sumarizadas as contribuições deste trabalho tanto para a organização em que este foi desenvolvido quanto para aplicações futuras por outros autores em contextos similares. Por fim, será discorrido sobre as limitações incidentes sobre o projeto e sobre as propostas de próximos passos necessários para um aproveitamento pleno do conteúdo deste trabalho pela organização.

6.1 Conclusões Gerais do Projeto

Para concluir o projeto, vale retomar seu objetivo: *"Consolidar, até o final do ano de 2022, planos de ação implementáveis que sejam capazes de alterar a distribuição dos perfis de cliente da base atual da Varejo ABC para uma composição que otimize a adesão desta base à proposta de valor da empresa"*. Ainda, é importante ressaltar que o projeto deveria atingir este objetivo de modo que estivesse alinhado com a realidade do mercado analisado na introdução do projeto, com a cultura da empresa e com os objetivos estratégicos desta.

Pode-se concluir que o projeto atingiu o objetivo projetado de acordo com os *feedbacks* recebidos de todos os *stakeholders* envolvidos no projeto. A clusterização elaborada no projeto já é de conhecimento de todas as lideranças da organização e já é, na data de conclusão deste trabalho, utilizada por diferentes áreas para moldar estratégias inerentes a suas rotinas. Além disso, o Painel de Controle de Potencial Máximo dos clientes também já é usado diariamente pela área de vendas da empresa para definir as prioridades dos vendedores dia a dia, buscando otimizar o tempo e os resultados deles.

A análise de transição dos clientes entre clusters é atualmente parte de uma iniciativa vigente na empresa para reestruturar a carteira de clientes geral da organização, assim como a distribuição ideal da base de clientes entre os clusters, que é o que serve de alvo para esta iniciativa. Os atributos correlacionados com a clusterização dos clientes ainda precisam ser melhor maturados dentro da organização e ainda não são utilizados efetivamente para o molde de personas para prospecção de clientes ou para eliminação de clientes com baixa correlação

com clusters de baixo potencial da base da empresa, mas já estão sendo melhor estudados por diversas pessoas na organização.

Dessa forma, é seguro afirmar que os resultados do projeto atenderam os objetivos esperados, sendo capazes de realizar um impacto significativo na rotina de diferentes áreas da empresa. Ainda, é possível dizer que os planos de ação a serem implementados futuramente se encontram em fase de priorização e de análise de esforço x impacto de cada um para a confirmação de sua aplicabilidade pela empresa.

6.2 Contribuição do Projeto para a Organização

Tendo em mente todos os aspectos mencionados acima, o projeto teve grande impacto nos indicadores de maior representatividade para a organização no âmbito atual: receita bruta, variedade de produtos vendidos e margem bruta. Com o impulsionamento destes indicadores pela iniciativa desenvolvida, ela também é fundamental para melhorar a rentabilidade da empresa. Retomam-se também os cinco fatores que motivaram a execução deste projeto que foram elencados na introdução.

Com relação à necessidade de redução do poder de barganha dos consumidores, o projeto permitiu à organização focar em clientes mais rentáveis para a empresa e que são mais direcionados a relações ganha-ganha com a Varejo ABC. Dessa forma, estes clientes tendem a ter maior adesão à proposta de valor da empresa e a serem mais fiéis à empresa, reduzindo assim o poder de barganha destes, haja vista que são clientes que veem valor em uma proposta de valor específica da empresa cliente.

No que tange à demanda por iniciativas para aumento das barreiras de entrada neste mercado para futuros concorrentes, pode-se dizer que a melhor exploração do potencial de compra dos clientes faz com que a empresa tenha maior representatividade nas compras realizadas por estas mercearias. Sendo assim, torna-se mais difícil que outras empresas substituam os serviços da organização nos clientes.

No aspecto da inevitabilidade da necessidade de realizar uma priorização de dimensões estratégicas a serem focadas neste mercado, pode-se afirmar que o projeto não foi conclusivo acerca da dimensão estratégica mais relevante para a organização. Porém, foi possível compreender que a empresa atende diferentes perfis de clientes, que por sua vez valorizam diferentes dimensões estratégicas. Com o direcionamento futuro da base de clientes a uma base que englobe mais clientes atrativos, será fundamental que sejam direcionadas

ações estratégicas para a consolidação das dimensões estratégicas de maior valor para esta nova composição de base de clientes.

Dada a imprescindibilidade de novas rodadas de investimento com alto capital levantado e baixa diluição de ações, ressalta-se que para o levantamento de uma rodada Series B com bons resultados a empresa precisa ter boa lucratividade. Este foi justamente um dos principais objetivos tratados ao longo de todo o projeto e é seguro afirmar que a transição do perfil atual de carteira de clientes para um perfil otimizado será de suma importância para a consolidação deste levantamento de capital de forma saudável para a organização.

Por fim, com relação à necessidade de se gerir bem os recursos financeiros da empresa no processo em que se são atingidos os outros quatro pilares, destaca-se que a melhoria da lucratividade de clientes tem benefícios também diretos nesta questão. Conforme a margem bruta percentual dos clientes, que foi um dos KPIs trabalhados no projeto aumenta, mais capital é liberado para a organização investir em suas operações e em inovações a fim de se destacar no mercado.

Tendo em conta o impacto do projeto em indicadores-chave para a estratégia da organização e nos cinco fatores que motivaram o projeto a ser realizado, pode-se afirmar que o projeto sustenta uma contribuição bastante significativa para a organização. É importante também destacar que, embora já satisfatório, o impacto do projeto atualmente ainda é pequeno perto do potencial impacto que este irá gerar futuramente com a implementação dos planos de ação recomendados pelo autor e com a transição a médio prazo da composição da base de clientes da empresa de forma efetiva após a execução destes planos de ação.

6.3 Contribuição do Projeto para a Ciência

Embora o projeto tenha sido focado na organização estudada, o projeto possui contribuições que extrapolam o âmbito da empresa em questão. A implementação de métodos de Machine Learning e de análises de perfil de clientes em empresas de varejo, sobretudo alimentar, ainda é bastante escassa na literatura acadêmica brasileira.

Sendo assim, embora o projeto não crie efetivamente conhecimento através de novos métodos para a academia, é possível afirmar que esta aplicação auxilia a materializar os conceitos utilizados para futuros autores. Com o entendimento deste trabalho, engenheiros de outras organizações deverão poder fazer uso dos conceitos aqui abordados para aplicação em sua realidade.

6.4 Limitações e Próximos Passos

Como todo projeto, as condições de desenvolvimento nem sempre são as desejadas pelo autor. No contexto deste projeto existem três limitações principais que tiveram que ser contornadas: o histórico limitado de compras apurável, a baixa quantidade de clientes com informações de perfil de loja mapeadas e a limitação de recursos humanos para implementação dos planos de ação do projeto.

De maio de 2022 para junho de 2022 houveram mudanças operacionais significativas na empresa, o que fez com que os dados anteriores a junho não fossem comparáveis com os dados posteriores a este mês. Dessa forma, o autor precisou trabalhar com um espaço amostral de 5 meses para elaborar sua clusterização. Caso houvesse acesso a período maiores de informações apuradas, a clusterização poderia ter obtido resultados ainda melhores, entretanto os resultados obtidos para este período já foram satisfatórios para os clientes.

Haveriam outras informações sobre o perfil de loja dos clientes que poderiam ser consultadas para um entendimento mais completo de como o perfil de cliente se correlaciona com seu perfil de compra determinado pela clusterização. Entretanto, o autor dispôs de uma quantidade limitada de informações sobre o perfil de loja dos clientes e um espaço amostral restrito de clientes que possuem essas informações mapeadas. Uma abrangência maior de dados referentes a este aspecto poderia ter contribuído para a identificação de características mais significativas para a clusterização dos clientes, facilitando o desenho de uma persona para cada cluster que tornasse mais didática a compreensão dos clusters.

O projeto originou diversos planos de ação bastante significativos para moldar a base de clientes da empresa. Naturalmente, seria benéfico que todos fossem implementados e que essa implementação ocorresse o quanto antes. Entretanto, os recursos humanos responsáveis pela implementação desses planos de ação nas diferentes áreas da empresa são limitados pois concorrem com outras iniciativas de melhorias estratégicas e com a manutenção das rotinas da empresa. Os planos de ação serão então priorizados e serão dados prioridade aos planos de ação com maior potencial de impacto na organização consumindo o menor esforço possível.

Os próximos passos do projeto estão relacionados justamente com a implementação dos planos de ação resultantes deste e com o acompanhamento desta implementação. Ao longo dos próximos meses o autor estará em contato próximo com o Business Owner da organização garantindo a priorização e a implementação dos planos de ação pela ordem de prioridade definida.

APÊNDICE A - APLICAÇÃO DO MÉTODO *K-MEANS* DE CLUSTERIZAÇÃO

No Apêndice A estão registrados todos os trechos de códigos necessários para a aplicação do método de clusterização *K-Means* para a segmentação dos clientes da base da empresa em grupos distintos. Estes códigos foram utilizados na composição dos resultados expostos por meio de gráficos e de tabelas no corpo do relatório.

Ilustração 15: Importação das bibliotecas necessárias para clusterização

1 - Import de bibliotecas e definição de constantes

In [0]:

```
# Import de tabelas
from merce.merce_model.tables import CONTRIBUTION_MARGIN_1_PER_LINE_ITEM, LINE_ITEMS
import mlflow
from merce import getLogger
from merce.spark import df_column_to_list

# Import de bibliotecas estruturais
import numpy as np
import pandas as pd
from pandas import DataFrame
from merce.imports import *
from typing import List, Optional
from pyspark.sql.types import StructType,StructField, StringType, IntegerType


# Import de bibliotecas de visualização
import matplotlib.pyplot as plt
import seaborn as sns
from mpl_toolkits.mplot3d import Axes3D
import plotly.graph_objects as go

# Import ferramentas para monitorar duração do modelo
import time
from datetime import date

# Import de ferramentas de padronização dos dados
from pyspark.ml.feature import MinMaxScaler
from pyspark.ml.feature import StandardScaler
from pyspark.ml.feature import RobustScaler

# Import do modelo KMeans
from sklearn import metrics
from sklearn.metrics import f1_score
from sklearn.metrics import pairwise_distances
# from sklearn.cluster import KMeans
from pyspark.ml import Pipeline
from pyspark.ml.clustering import KMeans
from pyspark.ml.feature import VectorAssembler
from pyspark.ml.functions import vector_to_array
from pyspark.ml.evaluation import ClusteringEvaluator

ANALYSIS_START_DATE = date(2022, 6, 1)
ANALYSIS_END_DATE = date(2022, 10, 31)
EXPERIMENT_NAME = "/Shared/models/customers_profile_clusterization"
EXPERIMENT_ID = '3806054703099307'
DEFAULT_MODEL_VERSION = 28
EXCEPTION_STORES = [
    'a2bdb45b-1781-47ca-9b86-bb4c91c33b70',
    'faeccf1f-73a7-47ca-b210-670a23da0c3a',
]
EXCEPTION_CUSTOMERS = [
    'd73a92e3-1e0a-4e52-a151-0e994ad35186'
]
```

Fonte: Elaboração Própria

Ilustração 16: Função *get_customers_kpis_dataset*

2 - Construção do Dataset

In [0]:
```
def get_customers_kpis_dataset(pandas_df: bool = False,
                               start_date: date = ANALYSIS_START_DATE,
                               end_date: date = ANALYSIS_END_DATE) -> SparkDF or DataFrame:
    stores = (
        LINE_ITEMS.read()
        .select("order_id", "product_id", "store_id")
    )
    dataset = (
        CONTRIBUTION_MARGIN_1_PER_LINE_ITEM.read()
        .join(stores, ["order_id", "product_id"], "inner")
        .filter((S.col("data_pedido_erp").between(start_date, end_date)) &
                (~S.col("store_id").isin(EXCEPTION_STORES)) &
                (~S.col("customer_store_id").isin(EXCEPTION_CUSTOMERS)))
        .withColumn("month", S.date_trunc('month', S.col("data_pedido_erp")))
        .groupBy("customer_store_id", "month")
        .agg(S.sum("receita_bruta").alias("receita_bruta"),
             S.countDistinct("product_id").alias("skus_distintos"),
             S.sum("margem_bruta").alias("margem_bruta"))
        .groupBy("customer_store_id")
        .agg(S.sum("receita_bruta").alias("receita_bruta"),
             S.avg("receita_bruta").alias("receita_bruta_media"),
             S.avg("skus_distintos").alias("media_de_skus_distintos"),
             S.avg("margem_bruta").alias("media_de_margem_bruta"),
             S.sum("margem_bruta").alias("margem_bruta"))
        .withColumn("margem_bruta_percentual_media", S.col("media_de_margem_bruta") / S.col("receita_bruta_media"))
        .withColumn("margem_bruta_percentual", S.col("margem_bruta") / S.col("receita_bruta"))
        .select("customer_store_id", "receita_bruta_media",
                "media_de_skus_distintos", "margem_bruta_percentual_media")
    )

    if pandas_df:
        dataset = dataset.toPandas().set_index('customer_store_id')

    return dataset
```

Fonte: Elaboração Própria

Ilustração 17: Função *get_assembled_pyspark_dataset*

4 - Aplicar Vector Assemble e Normalização para os dados contidos no dataset

In [0]:
```
def get_assembled_pyspark_dataset() -> SparkDF:
    spark_dataset = get_customers_kpis_dataset(False)
    input_columns = ["receita_bruta_media", "media_de_skus_distintos", "margem_bruta_percentual_media"]
    vector_assembler = VectorAssembler(inputCols=input_columns,
                                       outputCol="features")
    assembled_dataset = vector_assembler.transform(spark_dataset)
    return assembled_dataset
```

Fonte: Elaboração Própria

Ilustração 18: Função *get_standardized_pyspark_dataset*

```
In [0]:
def get_standardized_pyspark_dataset(standardization_method: str = 'RobustScaler') -> SparkDF:
    assembled_dataset = get_assembled_pyspark_dataset()
    if standardization_method == 'MinMaxScaler':
        min_max_scaler = MinMaxScaler(outputCol="scaled")
        min_max_scaler.setInputCol("features")
        min_max_scaler_model = min_max_scaler.fit(assembled_dataset)
        min_max_scaler_model.setOutputCol("features_scaled")
        standardized_pyspark_dataset = min_max_scaler_model.transform(assembled_dataset)
        return standardized_pyspark_dataset

    elif standardization_method == 'StandardScaler':
        standard_scaler = StandardScaler(outputCol="scaled")
        standard_scaler.setInputCol("features")
        standard_scaler_model = standard_scaler.fit(assembled_dataset)
        standard_scaler_model.setOutputCol("features_scaled")
        standardized_pyspark_dataset = standard_scaler_model.transform(assembled_dataset)
        return standardized_pyspark_dataset

    elif standardization_method == 'RobustScaler':
        standard_scaler = StandardScaler(outputCol="scaled")
        standard_scaler.setInputCol("features")
        standard_scaler_model = standard_scaler.fit(assembled_dataset)
        standard_scaler_model.setOutputCol("features_scaled")
        standardized_pyspark_dataset = standard_scaler_model.transform(assembled_dataset)
        return standardized_pyspark_dataset

    else:
        standardized_pyspark_dataset = assembled_dataset.withColumnRenamed("features", "features_scaled")
        return standardized_pyspark_dataset
```

Fonte: Elaboração Própria

Ilustração 19: Função *get_standardized_pandas_dataset*

```
In [0]:
def get_standardized_pandas_dataset(standardization_method: str = 'RobustScaler') -> DataFrame:
    standardized_spark_dataset = get_standardized_pyspark_dataset(standardization_method)
    standardized_spark_dataset = (
        standardized_spark_dataset
        .withColumn("array", vector_to_array("features_scaled"))
        .select(["customer_store_id"] + [S.col("array")[i] for i in range(3)])
        .withColumnRenamed("array[0]", "receita_bruta_media")
        .withColumnRenamed("array[1]", "media_de_skus_distintos")
        .withColumnRenamed("array[2]", "margem_bruta_percentual_media")
    )
    standardized_pandas_dataset = standardized_spark_dataset.toPandas()
    return standardized_pandas_dataset
```

Fonte: Elaboração Própria

Ilustração 20: Função *plot_elbow_method_chart*

6 - Definir os melhores valores para K, standardization_method e distance_measure

```
In [0]:
def plot_elbow_method_chart(standardization_method: str = 'RobustScaler',
                            distace_measure: str = 'euclidean') -> SparkDF:
    standardized_pyspark_dataset = get_standardized_pyspark_dataset(standardization_method)
    distortions = []
    K = range(2,11)
    for k in K:
        kmeans = KMeans(
            k=k, distanceMeasure=distace_measure, featuresCol="features_scaled",
            predictionCol="cluster_id", maxIter=1000
        )
        kmeans_fit = kmeans.fit(standardized_pyspark_dataset)
        distortion = kmeans_fit.summary.trainingCost
        distortions.append(distortion)
        print(f"Erro total para {k} clusters = {distortion}")

    plt.plot(K, distortions, 'bx-')
    plt.xlabel('Número de Clusters (k)')
    plt.ylabel('Erro Total do Método')
    plt.title(f"Método do Cotovelo destacando o valor ótimo de k utilizando {standardization_method} como normalizador e {distace_measure} como métric
    plt.show()
```

Fonte: Elaboração Própria

Ilustração 21: Função *plot_silhouette_score_chart*

```
In [0]:
def plot_silhouette_score_chart(standardization_method: str = 'RobustScaler',
                                distace_measure: str = 'euclidean') -> SparkDF:
    standardized_pyspark_dataset = get_standardized_pyspark_dataset(standardization_method)
    silhouette_scores=[]
    evaluator = ClusteringEvaluator(featuresCol='features_scaled',
                                    metricName='silhouette',
                                    distanceMeasure='squaredEuclidean')
    K = range(2,11)
    for k in K:
        kmeans = KMeans(
            k=k, distanceMeasure="euclidean",
            featuresCol="features_scaled", maxIter=1000
        )
        kmeans_fit = kmeans.fit(standardized_pyspark_dataset)
        kmeans_transform = kmeans_fit.transform(standardized_pyspark_dataset)
        evaluation_score = evaluator.evaluate(kmeans_transform)
        silhouette_scores.append(evaluation_score)
        print(f"Evaluation Score para {k} clusters = {evaluation_score}")

    plt.plot(K, silhouette_scores, 'bx-')
    plt.xlabel('Número de Clusters (k)')
    plt.ylabel('Silhouette Score')
    plt.title(f"Método Silhouette Score destacando o valor ótimo de k utilizando {standardization_method} como normalizador e {distace_measure} como m
    plt.show()
```

Fonte: Elaboração Própria

Ilustração 22: Função *train_kmeans_model_for_customer_profile_clusterization*

7 - Construir e gravar o modelo de Clusterização K-Means otimizado

```
In [0]:
def train_kmeans_model_for_customer_profile_clusterization(standardization_method: str,
                                                           n_clusters: int,
                                                           distace_measure: str) -> SparkDF:
    spark_dataset = get_customers_kpis_dataset(pandas_df=False)
    input_columns = ["receita_bruta_media", "media_de_skus_distintos", "margem_bruta_percentual_media"]

    with mlflow.start_run(experiment_id=EXPERIMENT_ID) as run:
        logger.info("run_id: {}".format(run.info.run_id))
        vector_assembler = VectorAssembler(inputCols=input_columns, outputCol="features")
        assembled_dataset = vector_assembler.transform(spark_dataset)

        if standardization_method == 'MinMaxScaler':
            min_max_scaler = MinMaxScaler(outputCol="scaled")
            min_max_scaler.setInputCol("features")
            scaler_model = min_max_scaler.fit(assembled_dataset)
            scaler_model.setOutputCol("features_scaled")

        elif standardization_method == 'StandardScaler':
            standard_scaler = MinMaxScaler(outputCol="scaled")
            standard_scaler.setInputCol("features")
            scaler_model = standard_scaler.fit(assembled_dataset)
            scaler_model.setOutputCol("features_scaled")

        elif standardization_method == 'RobustScaler':
            standard_scaler = RobustScaler(outputCol="scaled")
            standard_scaler.setInputCol("features")
            scaler_model = standard_scaler.fit(assembled_dataset)
            scaler_model.setOutputCol("features_scaled")

        if standardization_method == 'MinMaxScaler' or standardization_method == 'StandardScaler' or standardization_method == 'RobustScaler':
            kmeans = KMeans(
                k=n_clusters, distanceMeasure=distace_measure, featuresCol="features_scaled",
                predictionCol="cluster_id", maxIter=1000
            )
        else:
            kmeans = KMeans(
                k=n_clusters, distanceMeasure=distace_measure, featuresCol="features",
                predictionCol="cluster_id", maxIter=1000
            )

    if standardization_method == 'MinMaxScaler' or standardization_method == 'StandardScaler' or standardization_method == 'RobustScaler':
        pipeline = Pipeline(stages=[vector_assembler, scaler_model, kmeans])
    else:
        pipeline = Pipeline(stages=[vector_assembler,  kmeans])

    pipeline_model = pipeline.fit(spark_dataset)
    mlflow.spark.log_model(
        pipeline_model,
        artifact_path="models",
        registered_model_name="customers_profile_clustering"
    )
    mlflow.end_run()
```

Fonte: Elaboração Própria

Ilustração 23: Função *get_customers_clusterized_by_customer_profile*

In [0]:
```
def get_customers_clusterized_by_customer_profile(spark_dataset: DataFrame,
                                                   model_version: Optional[int] = None,
                                                   normalized: bool = False) -> SparkDF:
    if model_version is None:
        model_path = "models:/customers_profile_clustering/{}".format(DEFAULT_MODEL_VERSION)
    else:
        model_path = "models:/customers_profile_clustering/{}".format(model_version)
    model = mlflow.spark.load_model(model_path)
    mlflow.set_experiment(EXPERIMENT_NAME)
    with mlflow.start_run(experiment_id=EXPERIMENT_ID) as run:
        logger.info("run_id: {}".format(run.info.run_id))
        customers_clusterized_by_customer_profile = model.transform(spark_dataset)

        if normalized:
            customers_clusterized_by_customer_profile = (
                customers_clusterized_by_customer_profile
                .drop("receita_bruta_media", "media_de_skus_distintos", "margem_bruta_percentual_media")
                .withColumn("array", vector_to_array("features_scaled"))
                .select(["customer_store_id"] + [S.col("array")[i] for i in range(3)] + ['cluster_id'])
                .withColumnRenamed("array[0]", "receita_bruta_media")
                .withColumnRenamed("array[1]", "media_de_skus_distintos")
                .withColumnRenamed("array[2]", "margem_bruta_percentual_media")
                .select(["customer_store_id", "receita_bruta_media","media_de_skus_distintos",
                         "margem_bruta_percentual_media", "cluster_id"]).distinct()
            )

        else:
            customers_clusterized_by_customer_profile = (
                customers_clusterized_by_customer_profile
                .select(["customer_store_id", "receita_bruta_media","media_de_skus_distintos",
                         "margem_bruta_percentual_media", "cluster_id"]).distinct()
            )

        sizes = (
            df_column_to_list(
            customers_clusterized_by_customer_profile
            .groupBy("cluster_id")
            .agg(S.countDistinct("customer_store_id").alias("cluster_size"))
            .select("cluster_size")
            )
        )
        params = {
            "model": model_path,
            "n_clusters": len(sizes),
            "group": "customers",
            "clusters": " ".join(["{}:{}".format(cid, s) for cid, s in enumerate(sizes)]),
        }
        mlflow.log_params(params)
        mlflow.log_metric("min_cluster", min(sizes))
        mlflow.log_metric("avg_cluster", sum(sizes)/len(sizes))
        mlflow.log_metric("max_cluster", max(sizes))
    return customers_clusterized_by_customer_profile
```

Fonte: Elaboração Própria

Ilustração 24: Função *get_statistic_of_clusterized_pyspark_dataset*

In [0]:
```
def get_statistics_of_clusterized_pyspark_dataset(spark_dataset: SparkDF) -> SparkDF:
    statistics_of_clusterized_pyspark_dataset = spark_dataset.groupBy("cluster_id") \
        .agg(S.count("*").alias("number_of_customers"),
             S.round(S.avg("receita_bruta_media"), 2).alias("avg_receita_bruta_media"),
             S.round(S.stddev("receita_bruta_media"), 2).alias("stddev_receita_bruta_media"),
             S.round(S.min("receita_bruta_media"), 2).alias("min_receita_bruta_media"),
             S.round(S.max("receita_bruta_media"), 2).alias("max_receita_bruta_media"),
             S.round(S.avg("media_de_skus_distintos"), 2).alias("avg_media_de_skus_distintos"),
             S.round(S.stddev("media_de_skus_distintos"), 2).alias("stddev_media_de_skus_distintos"),
             S.round(S.min("media_de_skus_distintos"), 2).alias("min_media_de_skus_distintos"),
             S.round(S.max("media_de_skus_distintos"), 2).alias("max_media_de_skus_distintos"),
             S.round(S.avg("margem_bruta_percentual_media") * 100, 2).alias("avg_margem_bruta_percentual_media"),
             S.round(S.stddev("margem_bruta_percentual_media") * 100, 2).alias("stddev_margem_bruta_percentual_media"),
             S.round(S.min("margem_bruta_percentual_media") * 100, 2).alias("min_margem_bruta_percentual_media"),
             S.round(S.max("margem_bruta_percentual_media") * 100, 2).alias("max_margem_bruta_percentual_media")) \
        .orderBy("cluster_id")
    return statistics_of_clusterized_pyspark_dataset
```

Fonte: Elaboração Própria

Ilustração 25: Função *plot_statistics_of_clusterized_customers_pandas_dataset*

In [0]:

```
def plot_statistics_of_clusterized_customers_pandas_dataset(statistics_pandas_dataset: DataFrame) -> SparkDF:
    colors = ['red', 'blue', 'green', 'yellow', 'gray', 'purple', 'orange', 'pink']
    charts_axis = ['receita_bruta_media', 'media_de_skus_distintos', 'margem_bruta_percentual_media']
    for chart_axis in charts_axis:
        for cluster in np.unique(labels):
            x_axis_value = cluster
            y_axis_value = statistics_pandas_dataset[statistics_pandas_dataset['cluster_id'] == cluster][f"avg_{chart_axis}"].to_numpy()
            error = statistics_pandas_dataset[statistics_pandas_dataset['cluster_id'] == cluster][f"stddev_{chart_axis}"].to_numpy()
            plt.errorbar(x_axis_value,
                         y_axis_value,
                         error,
                         color=colors[cluster], label='Cluster ' + str(cluster), linestyle='None', marker='o')
            plt.annotate(statistics_pandas_dataset[statistics_pandas_dataset['cluster_id'] == cluster][f"avg_{chart_axis}"].to_numpy()[0],
                          (x_axis_value, y_axis_value[0] + 0.2))
        plt.xlabel('Cluster (k)')
        plt.ylabel(chart_axis)
        plt.title(f"Gráfico de média e desvio padrão de {chart_axis} por cluster")
        plt.legend()
        plt.show()
```

Fonte: Elaboração Própria

APÊNDICE B - PAINEL DE POTENCIAL MÁXIMO DE CLIENTES

Neste apêndice estão registrados os trechos de código em *SQL* produzidos na plataforma *Metabase* que foram necessários para a construção do Painel de Potencial Máximo de clientes produzido.

Ilustração 26: *Query SQL* para consulta de informações dos clientes

```sql
with customers as (
    select distinct customer_store_id, customer_store_name from contribution_margin_1_per_line_item
),

cluster as (
select
    customer_store_id,
    cluster_id
from
    curated.customer_profile_clusters
where
    month = '2022-10-01'
),
```

Fonte: Elaboração Própria

Ilustração 27: *Query SQL* para consulta de performance dos clientes no mês corrente

```sql
month_to_date as (
select
    customer_store_id,
    sum(receita_bruta) as receita_bruta_mtd,
    count(distinct product_id) as skus_distintos_mtd,
    sum(margem_bruta) / sum(receita_bruta) as margem_bruta_mtd
from
    contribution_margin_1_per_line_item
where
    data_pedido_erp between date_trunc('month', current_date - interval '1 day') and current_date - interval '1 day'
group by
    customer_store_id
),
```

Fonte: Elaboração Própria

Ilustração 28: *Query SQL* para consulta de Potencial Máximo dos Clientes

```
pmax as (
select
    customer_store_id,
    customer_store_name,
    max(receita_bruta) as receita_bruta_max,
    max(skus_distintos) as skus_distintos_max,
    max(margem_bruta / receita_bruta) as margem_bruta_max
from(
    select
        customer_store_id,
        customer_store_name,
        date_trunc('month', data_pedido_erp) as mes,
        sum(receita_bruta) as receita_bruta,
        count(distinct order_id) as number_of_orders,
        count(distinct product_id) as skus_distintos,
        sum(margem_bruta) as margem_bruta
    from
        contribution_margin_1_per_line_item
    group by
        customer_store_id, customer_store_name, mes) a
group by
    customer_store_id, customer_store_name
```

Fonte: Elaboração Própria

Ilustração 29: *Query SQL* do Painel de Controle de Clientes

```
select
    customer_store_id,
    receita_bruta_mtd,
    receita_bruta_max,
    receita_bruta_max - receita_bruta_mtd as delta_receita_bruta,
    skus_distintos_mtd,
    skus_distintos_max,
    skus_distintos_max - skus_distintos_mtd as delta_skus_distintos,
    margem_bruta_mtd,
    margem_bruta_max,
    margem_bruta_max - margem_bruta_mtd as delta_margem_bruta
from(
    select
        cs.customer_store_id,
        cs.customer_store_name,
        ifnull(receita_bruta_mtd, 0) as receita_bruta_mtd,
        ifnull(receita_bruta_max, 0) as receita_bruta_max,
        ifnull(skus_distintos_mtd, 0) as skus_distintos_mtd,
        ifnull(skus_distintos_max, 0) as skus_distintos_max,
        ifnull(margem_bruta_mtd, 0) as margem_bruta_mtd,
        ifnull(margem_bruta_max, 0) as margem_bruta_max
    from
        customers cs
        left join cluster c on cs.customer_store_id = c.customer_store_id
        left join month_to_date mtd on cs.customer_store_id = mtd.customer_store_id
        left join pmax pm on cs.customer_store_id = pm.customer_store_id
    [[where
        cluster_id = {{cluster}}]]) a
order by
    customer_store_name
```

Fonte: Elaboração Própria

APÊNDICE C - ANÁLISE DE TRANSIÇÃO DE CLIENTES ENTRE CLUSTERS

Para a construção dos gráficos de transição de clientes entre clusters, foram construídas as funções expostas neste apêndice. As ilustrações representam os códigos escritos em Python através de capturas de tela do repositório da empresa.

Ilustração 30: Função *get_transition_summary*

```
In [0]: def get_transition_summary(month_p1: date,
                                   month_p2: date,
                                   clusters_fonte: Optional[List[int]] = None,
                                   clusters_destino: Optional[List[int]] = None) -> SparkDF:
    clusterized_customers_p1 = (
        CUSTOMER_PROFILE_CLUSTERS.read()
        .filter(S.col("month") == month_p1)
        .withColumn("cluster_id_p1", S.when(S.col("cluster_id") == 'inativo', 7).otherwise(S.col("cluster_id")))
        .select("customer_store_id", "cluster_id_p1")
    )

    clusterized_customers_p2 = (
        CUSTOMER_PROFILE_CLUSTERS.read()
        .filter(S.col("month") == month_p2)
        .withColumn("cluster_id_p2", S.when(S.col("cluster_id") == 'inativo', 7).otherwise(S.col("cluster_id")))
        .select("customer_store_id", "cluster_id_p2")
    )

    clusters_transition = (
        clusterized_customers_p1
        .join(clusterized_customers_p2, "customer_store_id", "left")
    )

    transition_summary_base = (
        clusters_transition
        .groupBy("cluster_id_p1", "cluster_id_p2")
        .agg(S.countDistinct("customer_store_id").alias("customers"))
    )
    clusters_fonte = [0, 1, 2, 3, 4, 5, 6, 7]
    clusters_destino = [0, 1, 2, 3, 4, 5, 6, 7]

    data = []
    for cluster_fonte in clusters_fonte:
        for cluster_destino in clusters_destino:
            data_add = (cluster_fonte, cluster_destino)
            data.append(data_add)
```

```
    schema = StructType([ \
    StructField("cluster_id_p1", IntegerType(),True), \
    StructField("cluster_id_p2", IntegerType(),True), \
    ])

    dataset_base = spark.createDataFrame(data=data,schema=schema)

    transition_summary = (
        dataset_base
        .join(transition_summary_base, ["cluster_id_p1", "cluster_id_p2"], "left")
        .fillna(0)
        .withColumn("cluster_id_p2", (S.col("cluster_id_p2") + len(clusters_fonte)).cast(IntegerType()))
        .orderBy("cluster_id_p1", "cluster_id_p2")
    ).toPandas()

    return transition_summary
```

Fonte: Elaboração Própria

Ilustração 31: Função *plot_sankey_chart*

In [0]:

```
def plot_sankey_chart(transition_summary: SparkDF,
                      cluster_fonte: int) -> SparkDF:
    filtered_transition_summary = transition_summary[transition_summary['cluster_id_p1'] == cluster_fonte]
    colors = {
        0: "red",
        1: "blue",
        2: "green",
        3: "yellow",
        4: "gray",
        5: "purple",
        6: "orange",
        7: "black",
    }
    target = list(filtered_transition_summary["cluster_id_p2"])
    target_formated = [t - 7 for t in target]

    fig = go.Figure(data=[go.Sankey(
        node = dict(
          pad = 15,
          thickness = 20,
          line = dict(
              color = colors[cluster_fonte],
              width = 0.5
          ),
          label = ["C"+f"{cluster_fonte}", "C0", "C1", "C2", "C3", "C4", "C5", "C6", "Inativo"],
          color = colors[cluster_fonte]
        ),
        link = dict(
          source = [0] * len(target_formated),
          target = target_formated,
          value = list(filtered_transition_summary["customers"])
      ))])
    fig.update_layout(title_text=f"Transição de Clusters dos Clientes do Cluster {cluster_fonte} 3Q22 vs Últimos 30 Dias", font_size=10)
    fig.show()
```

Fonte: Elaboração Própria

REFERÊNCIAS BIBLIOGRÁFICAS

Instituto Brasileiro de Geografia e Estatística (2020). **Contas Nacionais Trimestrais**, Brasil, p.19, dez. 2020. Brasil Disponível em: <https://biblioteca.ibge.gov.br/visualizacao/periodicos/2121/cnt_2020_4tri.pdf> Acesso em Acesso em 27 mai. 2022.

Sociedade Brasileira de Varejo e Consumo (2021). **O Papel do Varejo na Economia Brasileira**. Disponível em: <http://sbvc.com.br/wp-content/uploads/2021/04/O-Papel-do-Varejo-na-Economia-Brasileira_2021-SBVC-4.pdf> Acesso em 27 mai. 2022.

Associação Brasileira de Supermercados (2021). **Ranking Abras 2021**: Os dados oficiais de um setor cada vez mais essencial, trazidos em parceria com a NielsenIQ. Super Hiper, Brasil, Ano 47, Número 537, Junho, 2021. Disponível em: <https://superhiper.abras.com.br/pdf/270.pdf> Acesso em 27 mai. 2022.

Sebrae (2015). **Pesquisa Minimercados no Brasil**. Serviço de Apoio às Micro e Pequenas Empresas. Disponível em: <https://bibliotecas.sebrae.com.br/chronus/ARQUIVOS_CHRONUS/bds/bds.nsf/3f5908f315baad2fb0ada9de370e4eaf/$File/5702.pdf> Acesso em 27 mai. 2022.

Alizila Staff (2018). **A Dose of New Retail for China´s Convenience Stores. Alizila News from Alibaba**, 2018. Disponível em: <https://www.alizila.com/alibaba-gives-dose-new-retail-china-convenience-stores/> Acesso em 28 de mai. 2022

COLLINS, David J.; RUKSTAD, Michael G. (2018). **Can You Say What Your Strategy Is?** Harvard Business Review, Reprint RO8O4E, 2018. Disponível em: <https://edisciplinas.usp.br/pluginfile.php/5430480/mod_resource/content/1/1.2%20Can%20you%20say%20what%20your%20strategy%20is%20-%20Collis%20and%20Rukstad%20%282008%29.pdf> Acesso em 28 de mai. 2022

OSTERWALDER, Alexander; PIGNEUR, Yves; BERNARDA, Gregory; SMITH, Alan. **Value Proposition Design: How to Create Products and Services Customers Want**. New Jersey, 2014. Disponível em: <https://books.google.com.br/books?id=jgu5BAAAQBAJ&printsec=frontcover&hl=pt-BR&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false> Acesso em 28 de mai. 2022

PORTER, Michael E. **Towards a Dynamic Theory of Strategy**. Strategic Management Journal, Vol. 12, p. 95-117, 1991. Disponível em: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/smj.4250121008> Acesso em 28 de mai. 2022

CARVALHO, Marly M.; LAURINDO, Fernando J. B. **Estratégia Competitiva: dos conceitos à implementação**. Editora Atlas, 2ª Edição. São Paulo, 2010. Disponível em: <https://edisciplinas.usp.br/pluginfile.php/6382542/mod_resource/content/1/Estrategia_Competitiva_dos_conceitos_a_i.pdf> Acesso em 28 de mai. 2022

DORAN, George T. **There´s a S.M.A.R.T. way to write management 's goals and objectives**. Management Review, Vol. 70, Issue 1, 1981. Disponível em: <https://community.mis.temple.edu/mis0855002fall2015/files/2015/10/S.M.A.R.T-Way-Management-Review.pdf> Acesso em 28 de mai. 2022

DOERR, John. **Avalie o que importa**. Tradução: Bruno Menezes. 1.ed. Rio de Janeiro: Alta Books, 2019. 305p. Acesso em 03 de ago. 2022

SCHEIN, Edgar H; SCHEIN, Peter. **Cultura Organizacional e Liderança**. 5.ed. São Paulo: Editora Atlas, 2009. 441p. Acesso em 03 de ago. 2022

RIES, Eric. **The Lean Startup: How Today's Entrepreneurs Use Continuous Innovation to Create Radically Successful Business**. 1.ed. United States Of America: Editora Atlas, 2011. 296. Acesso em 12 de ago. 2022

MINTO, Barbara. **The Minto Pyramid Principle: Logic in Writing, Thinking and Problem Solving**. United States Of America: Minto International Inc., 2010. Acesso em 28 de mai. 2022

UCLA (2021). **What is the difference between categorical, ordinal and interval variables?** Advanced Research Computing Statistics Methods and Data Analytics of University of California, 2021. Disponível em: <https://stats.oarc.ucla.edu/other/mult-pkg/whatstat/what-is-the-difference-between-categorical-ordinal-and-interval-variables/#:~:text=An%20ordinal%20variable%20is%20similar,low%2C%20medium%20and%20high> Acesso em 28 de mai. 2022

OHNO, Taiichi; BODEK, Norman. **Toyota Production System**, 1ª Edição. New York, 1988. Disponível em: <https://www.taylorfrancis.com/books/mono/10.4324/9780429273018/toyota-production-system-taiichi-ohno-norman-bodek> Acesso em 28 de mai. 2022

FERNANDO, J. Profit and Losses Statement Meaning, Importance, Types and Examples. Investopedia. Vancouver, 2022. Disponível em: <https://www.investopedia.com/terms/p/plstatement.asp>. Acesso em: 13 de out. 2022

COSTA, R. P. et al. Engenharia Econômica e Finanças. 1 ed. Elsevier São Paulo, 2009 cap.2 Disponível em: <https://edisciplinas.usp.br/pluginfile.php/5993987/mod_resource/content/3/02.pdf> Acesso em: 13 de out. de 2022.

DATABRICKS. **Databricks Documentation**. Azure Databricks Platform, 2022. Disponível em: <https://docs.databricks.com/> Acesso em 28 de mai. 2022

PYCHARM. **Learn PyCharm**. Jet Brains Learn, 2022. Disponível em: <https://www.jetbrains.com/pycharm/learn/> Acesso em 28 de mai. 2022

SPARK. **PySpark Documentation**. Apache Spark, 2022. Disponível em: <https://spark.apache.org/docs/latest/api/python/> Acesso em 28 de mai. 2022

ROSSUM, Guido Van. **PEP 8 - Style Guide for Python Code**. Python Enhancement Proposals, 2001. Disponível em: <https://peps.python.org/pep-0008/> Acesso em 28 de mai. 2022

POSTGRESQL. **PostgreSQL Documentation**. Postgresql, V 14.5. Disponível em: <https://www.postgresql.org/docs/current/> Acesso em 03 de set. 2022

METABASE. **Metabase Documentation**. Metabase, V 0.43. Disponível em: <https://www.metabase.com/docs/latest/> Acesso em 28 de mai. 2022

STARMER, Josh. **A Gentle Introduction to Machine Learning**, 2018. Disponível em: <https://www.youtube.com/watch?v=Gv9_4yMHFhI&list=PLblh5JKOoLUICTaGLRoHQDuF_7q2GfuJF&index=2> Acesso em 24 de jun. 2022

STARMER, Josh. **Machine Learning Fundamentals: Cross Validation**. StatQuest, 2018. Disponível em: <https://www.youtube.com/watch?v=fSytzGwwBVw> Acesso em 28 de mai. 2022

Portal Data Science. **Introdução à Clusterização e os Diferentes Método**s. Brasília, São Paulo, 2018. Disponível em: <https://portaldatascience.com/introducao-a-clusterizacao-e-os-diferentes-metodos/> Acesso em: 05 out. 2022

SUNDARAM, Gireesh. **Complete Guide to Clustering Techniques**. Tamil Nadu, 2020. Disponível em: <https://www.kaggle.com/code/gireeshs/complete-guide-to-clustering-techniques> Acesso em 05 de out. 2022

LIKAS, Aristidis; VLASSIS, Nikos; VERBEEK, Jakob J. **The global k-means clustering algorithm**. Pattern Recognition, Volume 36, Issue 2, 2003, Pages 451-461. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0031320302000602> Acesso em 28 de mai. 2022

STARMER, Josh. **StatQuest: K-means clustering**, 2020. Disponível em: <https://www.youtube.com/watch?v=4b5d3muPQmA&list=PLblh5JKOoLUICTaGLRoHQDuF_7q2GfuJF&index=39&t=23s> Acesso em 09 de set. 2022

JUNQUEIRA, L. **Estatística Descritiva: Análise Bidimensional**, 2019. Disponível em: <https://edisciplinas.usp.br/pluginfile.php/4863718/mod_resource/content/1/Aula-04_PRO3200-Estat%C3%ADstica.pdf> Acesso em: 22 de out. 2022