Segmentação Inteligente de Clientes Usando K-Means: Aplicação de Técnicas de Clusterização

Intelligent Customer Segmentation Using K-Means: Application of Clustering Techniques

Eliezer Pereira Guimarães
IFMG - Campus Ibirité
eliezerpereira1234@gmail.com

Lucas Henrique da Silva Pampoline
IFMG - Campus Ibirité
lspampoline@gmail.com

Marcos Prado Guimarães
IFMG - Campus Ibirité
marcospguimaraes02@gmail.com

Victor Henrique de Mattos Medeiros
IFMG - Campus Ibirité
victormattosgt@gmail.com

Rafael Lima Belem
IFMG - Campus Ibirité
rafaellimabelem@hotmail.com

Thiago Henrique Barbosa de Carvalho Tavares
IFMG - Campus Ibirité
thiago.tavares@ifmg.edu.br

RESUMO

Este artigo explora a segmentação inteligente de clientes utilizando o algoritmo K-Means, com foco em estratégias de marketing orientadas por dados. A base de dados utilizada contém informações detalhadas sobre 2000 clientes, incluindo variáveis como idade, renda, ocupação e educação. A metodologia emprega técnicas de pré-processamento, como normalização de dados e One-Hot Encoding, além do Método do Cotovelo para determinar o número ideal de clusters. A Análise de Componentes Principais (PCA) é utilizada para reduzir a dimensionalidade dos dados, melhorando a eficiência do K-Means. Os resultados demonstram melhorias na personalização de ofertas e na eficiência da alocação de recursos.

Palavras-chave

Segmentação de clientes, K-Means, clusterização, análise de dados, marketing, personalização, Python, PCA, normalização, One-Hot Encoding.

ABSTRACT

This paper explores intelligent customer segmentation using the K-Means algorithm, focusing on data-driven marketing strategies. The dataset contains detailed information on 2000 customers, including age, income, occupation, and education. The methodology employs preprocessing techniques such as data normalization, One-Hot Encoding, and the Elbow Method to determine the optimal number of clusters. Principal Component Analysis (PCA) reduces data dimensionality, enhancing K-Means efficiency. Results indicate improvements in offer personalization and resource allocation efficiency.

Keywords

Customer segmentation, K-Means, clustering, data analysis, marketing, personalization, Python, PCA, normalization, and One-Hot Encoding.

1. INTRODUÇÃO

A segmentação de clientes através de técnicas de clusterização desempenha um papel fundamental na gestão de negócios, particularmente em estratégias de marketing orientadas por dados. Em um cenário de mercado cada vez mais competitivo, o uso de dados para identificar padrões de comportamento entre os consumidores possibilita uma abordagem mais eficiente, alinhando produtos e serviços às necessidades específicas de cada grupo de clientes. Ferramentas como K-Means, K-Medoids e métodos híbridos têm sido amplamente utilizadas para otimizar a segmentação, proporcionando às empresas uma visão clara sobre como melhor atender e fidelizar diferentes segmentos de clientes (JAIN, 2010).

A clusterização vai além da segmentação tradicional baseada em fatores demográficos ou geográficos. Ela permite uma análise mais detalhada do comportamento do consumidor, incorporando variáveis como preferências de compra, interações digitais e respostas a campanhas anteriores. Este aprofundamento possibilita uma segmentação mais eficaz, focada em comportamentos e padrões de consumo, ao invés de simples agrupamentos demográficos. Por exemplo, Patro e Sahu (2015) explicam que a normalização dos dados e o pré-processamento adequado são fundamentais para que técnicas de clusterização como o K-Means identifiquem corretamente os padrões significativos de consumo e preferências.

Além de melhorar a personalização de ofertas, a clusterização também é essencial para a otimização de recursos dentro das empresas. Ao identificar os grupos de clientes mais lucrativos ou os que demandam maior atenção, as organizações podem alocar seus recursos de forma mais eficiente, concentrando esforços em áreas com maior potencial de retorno. Segundo MacQueen (1967), o algoritmo K-Means é particularmente útil para criar perfis de consumidores em ambientes com grandes volumes de dados, o que permite o desenvolvimento de estratégias de marketing mais direcionadas e eficientes.

Um exemplo de aplicação da clusterização é no e-commerce, onde a coleta de dados comportamentais permite uma compreensão mais profunda das preferências dos consumidores. Grandes plataformas de comércio eletrônico, como a Amazon, utilizam algoritmos de clusterização para personalizar recomendações, ajustar a interface de navegação e até prever a demanda por produtos. Zheng e Casari (2018) destacam que essa personalização é essencial para melhorar a experiência do cliente e aumentar o valor médio das transações, pois as recomendações baseadas em dados comportamentais têm maior probabilidade de conversão.

Além disso, técnicas de clusterização têm sido empregadas em estratégias de retenção e fidelização de clientes. Através da análise preditiva, as empresas conseguem identificar padrões de comportamento que indicam uma possível deserção de clientes, o que permite que medidas preventivas sejam adotadas de forma proativa. Por exemplo, Jolliffe e Cadima (2016) sugerem que a redução de dimensionalidade usando Análise de Componentes Principais (PCA) pode auxiliar na identificação das variáveis mais influentes que indicam comportamentos de risco de cancelamento, melhorando a eficácia das estratégias de retenção.

A análise de dados em tempo real também se tornou uma prática comum em campanhas de marketing baseadas em eventos. Ferramentas como cookies permitem que os websites capturem dados detalhados sobre o comportamento dos visitantes, possibilitando a criação de ofertas personalizadas em tempo real. Segundo Kraemer e Blasey (2004), a correlação adequada entre variáveis pode ser usada para identificar preferências dos consumidores durante a navegação, mostrando um aumento significativo nas taxas de conversão e engajamento dos usuários.

No entanto, o uso de algoritmos de clusterização, como o K-Means, apresenta desafios, como a escolha do número adequado de clusters e a sensibilidade à escolha dos centróides iniciais. Kodinariya e Makwana (2013) abordam o método do cotovelo como uma técnica eficaz para determinar o número ideal de clusters, enquanto outras estratégias, como o uso de técnicas híbridas, podem resultar em maior precisão na identificação dos clusters e maior estabilidade nos resultados.

Outra técnica complementar à clusterização é a Análise de Componentes Principais (PCA), que, ao reduzir a dimensionalidade dos dados, facilita a visualização e interpretação dos clusters. Jolliffe e Cadima (2016) destacam que, embora o PCA não seja um método de clusterização em si, ele auxilia na identificação das variáveis mais influentes e na eliminação do ruído nos dados, melhorando a eficiência dos algoritmos de clusterização.

Sendo assim, a aplicação de técnicas de clusterização em combinação com métodos de análise preditiva e redução de dimensionalidade não apenas melhora a precisão da segmentação de clientes, como também oferece uma vantagem competitiva para empresas que buscam maximizar seu desempenho no mercado. Ao utilizar algoritmos de clusterização, as empresas podem desenvolver estratégias personalizadas e adaptáveis, capazes de atender às dinâmicas necessidades de seus consumidores, ao mesmo tempo em que otimizam suas operações e aumentam a rentabilidade.

2. LEVANTAMENTO BIBLIOGRÁFICO

O agrupamento de clientes por meio de técnicas de segmentação desempenha um papel essencial no marketing de serviços contemporâneo, pois facilita a personalização de estratégias e aprimora a relação entre empresas e consumidores. Essa prática permite que as organizações agrupem clientes com características semelhantes, o que as ajuda a identificar as necessidades e comportamentos de diferentes grupos, tornando possível a criação de campanhas e ofertas mais direcionadas. Além disso, a segmentação permite que as empresas alinhem seus produtos e serviços às expectativas de seus clientes, maximizando assim a relevância e a eficácia de suas iniciativas de marketing.

As empresas se beneficiam do agrupamento de clientes de várias formas. Primeiramente, a segmentação permite a identificação de oportunidades de mercado mais precisas, facilitando a previsão de tendências e a personalização de ofertas com base em eventos e gatilhos específicos. “Isso não apenas ajuda a melhorar a experiência do cliente, mas também aumenta as taxas de conversão” (BAILEY et al., 2009), uma vez que as propostas se tornam mais relevantes e alinhadas às necessidades dos consumidores. Um exemplo prático desse benefício é o uso de modelos de propensão, que ajudam as empresas a prever o comportamento dos clientes e identificar o momento mais oportuno para oferecer produtos ou serviços. Essas análises, baseadas em eventos significativos na vida do cliente, como a expiração de um contrato ou uma mudança de endereço, são particularmente úteis para aprimorar a comunicação personalizada, como tratado neste artigo.

Outro aspecto importante do agrupamento de clientes é sua contribuição para a retenção e fidelização. Ao compreender melhor o comportamento dos consumidores, as empresas podem identificar sinais de insatisfação ou possíveis deserções, o que lhes permite adotar medidas proativas para evitar a perda de clientes. Por exemplo, o estudo de eventos e gatilhos na jornada do consumidor pode indicar momentos em que um cliente está mais propenso a trocar de fornecedor ou quando uma ação corretiva, como a oferta de um benefício exclusivo, pode ser mais eficaz (BAILEY et al., 2009).

Além disso, a segmentação não se limita apenas à diferenciação de clientes por características demográficas ou geográficas. Cada vez mais, as empresas estão implementando estratégias de segmentação baseadas em comportamento e necessidades. Esse foco na individualização é crucial para garantir que cada cliente receba uma experiência personalizada e relevante, o que, por sua vez, aumenta o engajamento e a satisfação. As empresas que utilizam essa abordagem se tornam mais ágeis e capazes de responder rapidamente às mudanças nas expectativas dos clientes, o que lhes confere uma vantagem competitiva significativa.

Portanto, o agrupamento de clientes, apoiado por uma análise detalhada de dados e eventos, continua sendo uma ferramenta estratégica poderosa para as empresas que buscam melhorar sua performance no mercado. Ao integrar essas práticas em suas operações, as empresas podem não apenas atender melhor às necessidades de seus clientes, mas também otimizar suas operações e maximizar a rentabilidade de suas ações (BAILEY et al., 2009).

A segmentação de clientes utilizando técnicas de clusterização tem se mostrado uma ferramenta essencial no contexto empresarial contemporâneo. Em um mercado cada vez mais competitivo,

compreender os diferentes perfis de consumidores é fundamental para otimizar as estratégias de marketing, personalizar ofertas e melhorar a alocação de recursos. A clusterização, por meio de algoritmos que identificam padrões e características comuns entre grupos de clientes, possibilita uma abordagem mais eficiente e focada nas necessidades específicas de cada segmento (REDDY et al., 2023).

A qualidade dos sites, por exemplo, desempenha um papel crucial nesse processo, e os cookies são ferramentas essenciais para aprimorar essa qualidade. Ao armazenar pequenos arquivos de texto no dispositivo do usuário, os cookies permitem que os sites coletem dados detalhados sobre o comportamento dos visitantes, como páginas visitadas, tempo gasto e interações realizadas (Kuan et. al., 2008).

Essas informações são fundamentais para entender melhor as preferências dos clientes e personalizar suas experiências. Com dados sobre produtos visualizados e categorias de interesse, as empresas podem oferecer recomendações precisas e ajustar estratégias de marketing para atender às necessidades dos clientes de maneira mais eficaz. Isso resulta em uma maior taxa de conversão, pois sites que oferecem uma navegação mais fluida e um conteúdo relevante tendem a incentivar mais compras.

Além disso, os cookies também são cruciais para a retenção de clientes. Eles possibilitam uma experiência contínua e personalizada, lembrando das preferências e interações passadas dos usuários. Com essa personalização, os clientes se sentem mais valorizados, o que pode aumentar sua lealdade e a probabilidade de compras repetidas. Por exemplo, sites podem sugerir produtos complementares com base no histórico de compras ou simplificar o processo de login para facilitar o acesso em visitas futuras.

Historicamente, o K-Means, proposto por J.B. MacQueen em (LEE et. al., 1980), foi um dos primeiros algoritmos amplamente utilizados para a clusterização de dados. Este método se baseia na divisão de um conjunto de dados em K grupos ou clusters, onde cada ponto é atribuído ao cluster mais próximo com base na distância euclidiana. O K-Means destaca-se pela simplicidade e eficiência, tornando-se uma escolha popular em áreas como mineração de dados e reconhecimento de padrões. No entanto, esse algoritmo apresenta algumas limitações, sendo a principal a sua dependência na escolha dos pontos focais iniciais, que podem influenciar significativamente os resultados obtidos. Entretanto, um desafio constante no uso do K-Means é a determinação do número ideal de clusters (KASHWAN; VELU, 2013).

Com o tempo, surgiram melhorias para superar essas deficiências. Um exemplo é o desenvolvimento de algoritmos híbridos, que combinam o K-Means com outras técnicas. Um avanço importante descrito na literatura é a combinação do algoritmo de maior distância mínima com o K-Means tradicional, resultando em uma versão aprimorada do K-Means. Esta abordagem visa resolver dois dos principais problemas do K-Means tradicional: a dependência excessiva da escolha dos pontos focais iniciais e a tendência de ficar preso em mínimos locais. O algoritmo melhorado utiliza a distância mínima para escolher os pontos focais iniciais de forma mais inteligente, garantindo que eles estejam mais distantes entre si e sejam mais representativos. Como resultado, a precisão, velocidade e estabilidade do agrupamento são significativamente aumentadas, conforme demonstrado em experimentos comparativos.

Além do K-Means, o K-Medoids também é utilizado em processos de clusterização, principalmente por sua robustez a outliers. Ao contrário do K-Means, que calcula a média dos pontos para determinar os centróides, o K-Medoids utiliza objetos reais como representantes de clusters, reduzindo a influência de dados extremos que poderiam distorcer os resultados. Dessa forma, essa técnica oferece uma segmentação mais precisa em cenários onde há grande variação entre os clientes (REDDY et al., 2023).

Outro recurso frequentemente associado à clusterização é a Análise de Componentes Principais (PCA), que auxilia na redução da dimensionalidade dos dados. Embora não seja uma técnica de clusterização em si, o PCA facilita a visualização dos grupos gerados ao condensar as informações mais relevantes dos dados em um menor número de variáveis. Essa simplificação permite uma interpretação mais clara dos clusters e das características predominantes em cada grupo (KASHWAN; VELU, 2013).

A aplicação dessas técnicas de clusterização não apenas melhora a precisão na segmentação de clientes, mas também traz benefícios práticos para as empresas. A partir de uma compreensão mais detalhada dos perfis dos consumidores, torna-se possível personalizar campanhas de marketing, desenvolver produtos mais alinhados às expectativas de cada grupo e aprimorar o atendimento ao cliente. "Em última instância, a clusterização contribui para o aumento da competitividade empresarial, permitindo que as empresas respondam de forma mais ágil e assertiva às demandas de um mercado dinâmico" (REDDY et al., 2023, p. 363).

Portanto, a clusterização se configura como uma prática indispensável para empresas que buscam maximizar seu desempenho no mercado. A utilização de técnicas como K-Means, K-Medoids e PCA possibilita uma segmentação eficiente e baseada em dados, proporcionando insights valiosos sobre o comportamento dos clientes e orientando decisões estratégicas que impulsionam o crescimento e a inovação no ambiente corporativo (KASHWAN; VELU, 2013).

3. METODOLOGIA

3.1 Base de Dados

A base de dados utilizada para a realização deste estudo foi a Segmentation Data, disponível na Kaggle, que é uma plataforma de aprendizado de ciência de dados do Google. Esta base de dados contém informações detalhadas sobre 2000 clientes, sendo essas variáveis fundamentais para entender diferentes segmentos de clientes e comportamentos de consumo.

Essa base de dados se torna muito útil para direcionar estratégias de marketing, promoções específicas e melhorar o relacionamento com os clientes, tornando os processos de decisão mais eficientes. A tabela abaixo apresenta as variáveis presentes no banco de dados, assim como o tipo de dado, sendo ele é numérico ou categórico e uma descrição do que é cada um dos valores representados.

Variável	*Tipo de Dados*	*Intervalo*	*Descrição*
ID	Numérico	Inteiro	Identificador único de um cliente.
Sexo	Categórico	{0,1}	Sexo biológico do cliente: 0 - masculino, 1 - feminino.
Estado civil	Categórico	{0,1}	Estado civil do cliente: 0 - solteiro, 1 - não solteiro.
Idade	Numérico	Inteiro (18-76)	Idade do cliente em anos.
Educação	Categórico	{0,1,2,3}	Nível de educação: 0 - outro/desconhecido, 1 - ensino médio, 2 - universidade, 3 - pós-graduação.
Renda	Numérico	Real (35832-309364)	Renda anual auto-relatada em dólares.
Ocupação	Categórico	{0,1,2}	Categoria de ocupação: 0 - desempregado/não qualificado, 1 - empregado qualificado, 2 -gestão/autônomo.
Tamanho do assentamento	Categórico	{0,1,2}	Tamanho da cidade: 0 - pequena, 1 - média, 2 - grande.

Figura 1 - Fonte: Próprios Autores

3.2 Ferramentas e Bibliotecas

Para a implementação da clusterização na base de dados Clientes Mercado, foi utilizado a linguagem de programação Python, por se tratar de uma linguagem orientada a objetos amplamente utilizada para estruturar, tratar, analisar e aplicar técnicas de inteligência artificial em base de dados (LUZ, DO da et al., 2023).

A biblioteca Pandas foi essencial para a importação e manipulação dos dados, permitindo uma análise detalhada e a verificação de tipos de dados, dados faltantes e estatísticas descritivas. Junto a isso, o NumPy forneceu suporte para operações numéricas eficientes, crucial para o tratamento e manipulação de grandes conjuntos de dados.

Para a visualização de dados, Matplotlib e Seaborn foram utilizadas, permitindo a criação de gráficos que facilitam a interpretação e análise visual dos resultados.

A biblioteca scikit-learn foi utilizada para implementar técnicas como o One-Hot Encoding, que transforma variáveis categóricas em formatos binários adequados para algoritmos de machine learning. Além disso, a biblioteca facilitou a normalização de valores, ajustando os dados para uma escala comum, e a aplicação do algoritmo de K-Means.

O Método do Cotovelo foi empregado para determinar o número ideal de clusters, garantindo a eficiência da segmentação. Essas ferramentas permitiram a construção de um modelo eficiente e a condução de análises indispensáveis. A análise de correlação de valores também foi conduzida para identificar relações significativas entre variáveis, utilizando a biblioteca Plotly para criar mapas de calor interativos.

A técnica de PCA (Principal Component Analysis) foi aplicada para reduzir a dimensionalidade dos dados, preservando a variância essencial. Essa etapa foi crucial para melhorar a performance dos modelos de inteligência artificial e facilitar a visualização dos clusters formados durante a manipulação dos dados dos clientes.

3.3 Correlação de Valores:

A correlação entre variáveis numéricas foi calculada para identificar relações significativas que podem influenciar a segmentação dos clientes. Essa análise permite entender como diferentes atributos interagem, auxiliando na escolha das características mais relevantes para o modelo. Conforme destaca García, Luengo e Herrera (2015), "A identificação de correlações entre variáveis é um processo crítico para selecionar características relevantes e melhorar o desempenho dos algoritmos de aprendizado de máquina".

3.4 Normalização de Dados:

A normalização dos dados é fundamental para algoritmos como o K-Means, que são sensíveis à escala das variáveis. García, Luengo, e Herrera (2015) afirmam que "o ajuste da escala das variáveis garante que nenhuma característica domine a formação dos clusters, promovendo uma análise mais equilibrada". Para a implementação no artigo, utilizamos a normalização MinMaxScaler para ajustar os valores entre 0 e 1, com o intuito de garantir que todas as variáveis tenham um impacto equilibrado na análise.

3.5 One Hot Encoder:

One Hot Encoding é uma técnica fundamental para a transformação de variáveis categóricas em um formato numérico que possa ser utilizado em algoritmos de aprendizado de máquina. Segundo Zheng e Casari (2018), "A codificação adequada de variáveis categóricas é crucial para que os modelos de aprendizado de máquina processem os dados corretamente, sem criar uma falsa hierarquia ou relação entre as categorias". Na programação, essa técnica foi utilizada para transformar as variáveis categóricas (sexo, estado civil, educação, etc) em elementos numéricos, permitindo assim que essas informações fossem incluídas na análise de forma adequada.

3.6 Método do Cotovelo:

Para determinar o número ideal de clusters no K-Means, utilizamos o método do cotovelo. Este método consiste em calcular o "Within-Cluster Sum of Squares" (WCSS) para diferentes números de clusters e, em seguida, criar um gráfico com o número de clusters no eixo x e o WCSS no eixo y. O ponto de inflexão do gráfico (o "cotovelo") indica o número ótimo de clusters. Ketchen & Shook (1996) afirmam em seu trabalho que "A aplicação do método do cotovelo ajuda a visualizar o ponto em que a inclusão de novos clusters deixa de trazer melhorias significativas à variância explicada".

3.7 Método PCA:

A Análise de Componentes Principais (PCA) é uma técnica utilizada para simplificar conjuntos de dados com muitas variáveis, reduzindo a sua dimensionalidade enquanto mantém a maior parte da variabilidade original. Em bases de dados com múltiplas variáveis, é comum haver redundância de informações, o que pode dificultar a interpretação dos resultados dos algoritmos de aprendizado. Segundo Wold et al. (1987), "o PCA é essencial para simplificar a análise de dados complexos, destacando os padrões mais significativos". Neste trabalho, o PCA foi empregado para transformar as variáveis originais em componentes principais não

correlacionadas, escolhendo aqueles que melhor representam a variância do conjunto de dados.

A variância explicada “explained variance” no PCA (Análise de Componentes Principais) refere-se à quantidade de variabilidade nos dados originais que é capturada por cada componente principal. Cada componente principal é uma combinação linear das variáveis originais, ordenada de forma que a maior quantidade de variância possível seja capturada no primeiro componente, a segunda maior no segundo componente, e assim por diante. Jolliffe & Cadima (2016) recomendam "manter componentes que expliquem entre 70% e 90% da variância original para garantir uma representação adequada dos dados, evitando ruído desnecessário".

3.8 K-means:

Após normalizar os dados, codificar e reduzir as componentes principais, aplicamos o algoritmo K-Means para a formação dos clusters. O K-Means, conforme MacQueen (1967), "É uma técnica eficiente para agrupar dados minimizando a variância interna dos clusters, o que facilita a identificação de padrões e a compreensão das características comuns em cada grupo". Isso o torna adequado para descobrir perfis ocultos de clientes e oferecer insights valiosos para a tomada de decisão.

O K-means inicia com a seleção aleatória de K centroides, que representam os centros iniciais dos clusters. Em seguida, cada ponto de dados é atribuído ao cluster cujo centroide está mais próximo, geralmente medido pela distância euclidiana. Após a atribuição, os centroides são recalculados como a média de todos os pontos em cada cluster. Esse processo itera até que os centroides se estabilizam ou um critério de parada seja atingido.

A aplicação do algoritmo K-Means justifica-se pela necessidade de segmentar os clientes em grupos com características semelhantes, auxiliando na identificação de padrões e na elaboração de estratégias direcionadas de marketing

4. DESENVOLVIMENTO

No desenvolvimento do código de clusterização inicialmente foi realizada a importação da base de dados com a utilização da biblioteca pandas. Em seguida, com as ferramentas dessa biblioteca, foi realizada uma análise detalhada dos dados presentes nas colunas, sendo esta, a verificação do tipo de dado de cada coluna, a identificação de dados faltantes ou nulos, e a análise da distribuição dos dados. A implementação dessas ferramentas no código foi realizada por meio do método .info() que nos retorna os tipos de dados de cada coluna, o método .describe() que retorna informações estatísticas detalhadas dos dados e o método .isnull() onde por meio do mesmo foi verificado a existência da dados nulos nas colunas.

Em seguida, foi realizada a análise das distribuições das variáveis numéricas de Idade e Renda. Para isso, foram implementados gráficos de distribuição por meio de utilização da biblioteca seaborn, onde é possível visualizar a Figura 1 a distribuição dos dados de idade dos clientes, sendo o eixo x do gráfico a idade e o eixo y a frequência em que a idade aparece na base de dados, onde é possível observar por meio do gráfico que a a maior parte dos clientes presentes na base de dados possuem entre 20 e 30 anos. Na Figura 2 temos a distribuição de renda dos clientes sendo o eixo x a renda e o eixo y a frequência em que a renda aparece a na base dados, onde é possível observar que os clientes em sua maioria possuem renda entre US$100000 e US$150000 por ano. Esses gráficos fornecem uma visão detalhada sobre como as Idade e a Renda estão distribuídas, ajudando a entender a variabilidade e a forma dos dados.

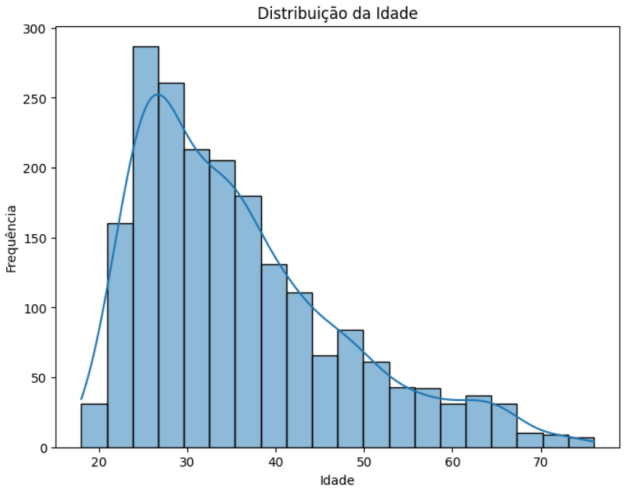


Figura 2 - Fonte: Próprios Autores

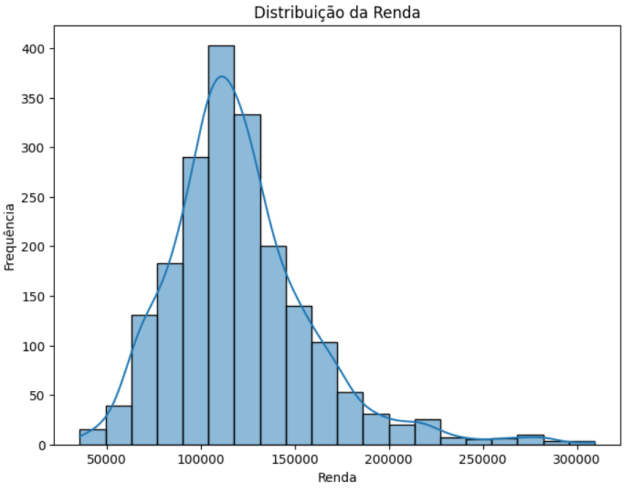


Figura 3 - Fonte: Próprios Autores

Para as variáveis categóricas de Sexo, Estado Civil, Educação e Tamanho da Cidade, foi realizada uma análise de distribuição utilizando gráficos de de frequencia por categoria, sendo utilizado a biblioteca seaborn para a plotagem dos graficos referentes a esses dados. Na Figura 3 temos a distribuição de de sexo por categoria sendo a categoria 0 sexo masculino e categoria 1 sexo feminino, onde é possivel observar que a distribuição de gêneros dos clientes são bastante equilibrados.

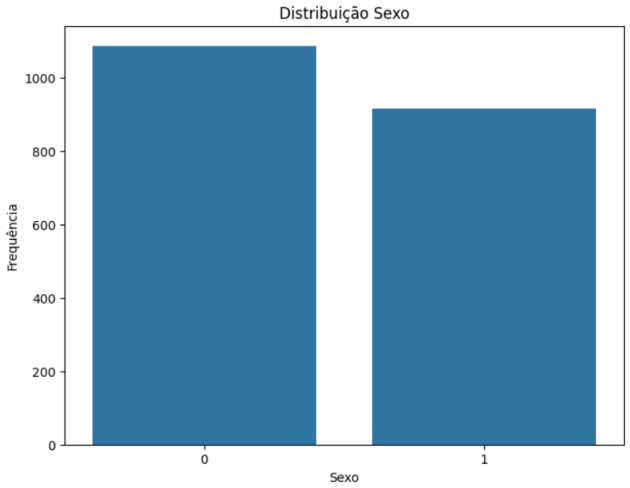


Figura 4 - Fonte: Próprios Autores

Para a distribuição de dados da coluna de educação, temos as categorias 0, 1, 2 e 3, sendo essas categorias correspondendo respectivamente a outro/desconhecido, ensino médio, graduação e pós-graduação. Por meio da Figura 4 é possível observar que na distribuição desses dados para os clientes temos uma concentração do nível de escolaridade no ensino médio.

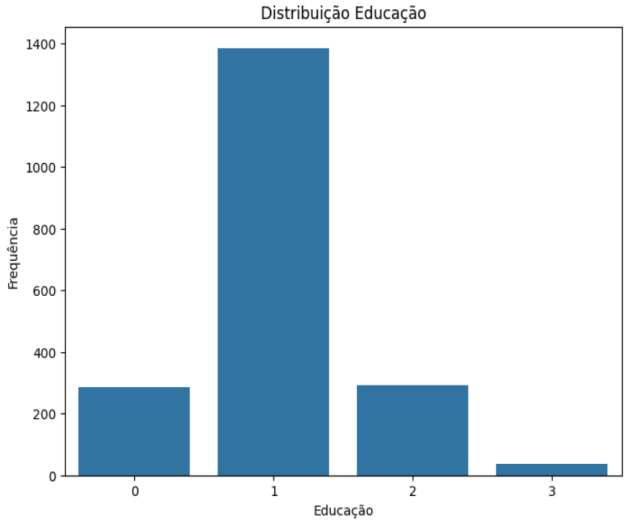


Figura 5 - Fonte: Próprios Autores

Na distribuição de dados da coluna de Estado Civil, a mesma possui as categorias 0 e 1 sendo que a categoria 0 é referente a clientes solteiros e a categoria 1 a clientes não solteiros, podendo os mesmos serem divorciados, separados, casados ou viúvos. Na Figura 5 temos a distribuição dos dados de estado civil onde por meio do gráfico é possível visualizar que a distribuição entre as duas categorias apresenta um equilíbrio.

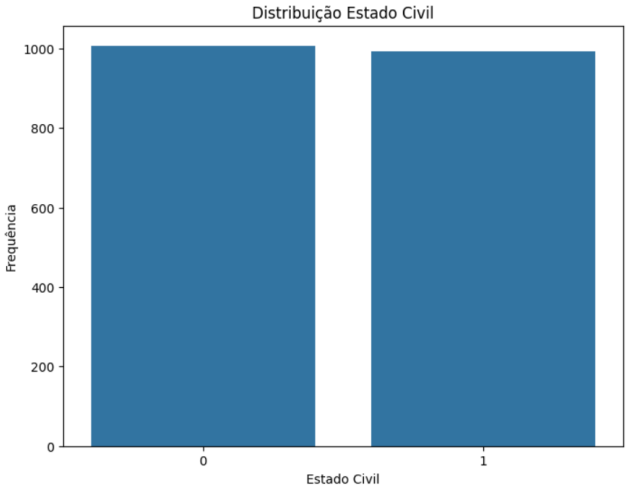


Figura 6 - Fonte: Próprios Autores

A coluna de dados de Ocupação é dividida em 3 categorias distintas, sendo elas 0, 1 e 2 onde elas apresentam os dados respectivamente de desempregado, empregado qualificado e empregado altamente qualificado. Na Figura 5 temos a distribuição dos dados de ocupação, sendo possível observar que os clientes em sua maioria são empregados qualificados.

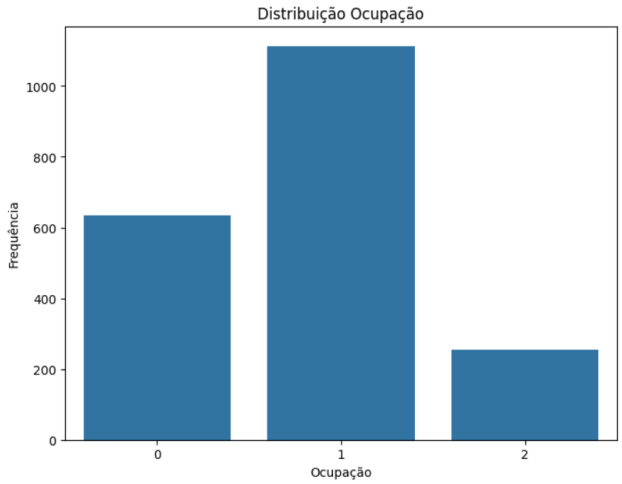


Figura 7 - Fonte: Próprios Autores

A coluna de dados de Tamanho da Cidade, nos traz os dados referentes ao tamanho da cidade onde o cliente reside sendo esta distribuída em 3 categorias distintas a 0, 1 e 2 as quais correspondem respectivamente a cidade pequena, cidade média e cidade grande. Na Figura 6 temos a distribuição dos dados referente ao tamanho da cidade em que os clientes residem sendo que os mesmo em sua maioria residem em cidades pequenas.

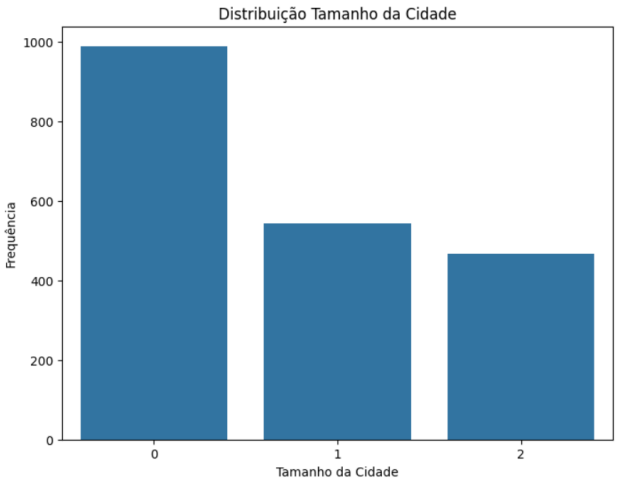


Figura 8 - Fonte: Próprios Autores

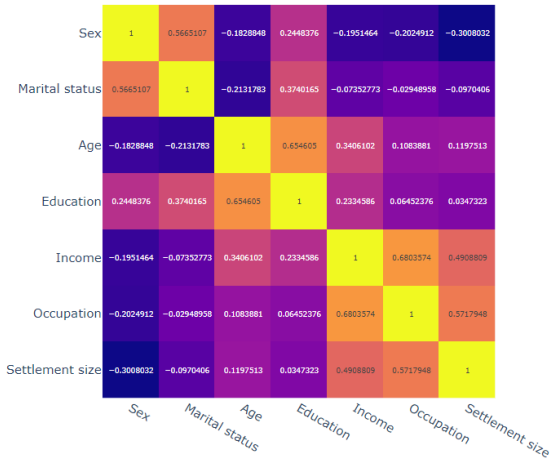


Figura 9 - Fonte: Próprios Autores

Em seguida, foi realizada a análise de relação entre variáveis sendo uma etapa fundamental pois por meio da mesma é possível identificar as variáveis do conjunto de dados que possuem relação entre si. A correlação entre variáveis pode indicar como uma variável pode influenciar ou estar associada a outra, sendo essencial para a seleção de features relevantes e para a interpretação de padrões nos dados. Para a implementação e plotagem do gráfico da correlação em Python, foi utilizado a biblioteca plotly.express, a qual possui a função px.imshow(), sendo que por meio da mesma foi gerado um mapa de calor das correlações, onde cada célula do gráfico representa a correlação entre pares de variáveis. Na Figura 7 é possível ver a matriz de correlação dos dados utilizados onde nos mostra que algumas variáveis possuem correlações moderadas a fortes, como "Educação" e "Idade" (0.654605) e de "Ocupação" e "Renda" (0.680374), o que sugere uma relação direta entre essas características. Já variáveis como "Sexo" e "Tamanho da Cidade" apresentam uma correlação negativa (-0.3008032), o que pode indicar que essas variáveis não possuem uma relação inversa em alguns aspectos. Com base nessa análise, não foi necessário remover variáveis para a implementação do código de inteligência artificial, uma vez que, apesar de algumas variáveis apresentarem correlação, ela não é suficientemente alta a ponto de justificar a exclusão das mesmas.

O próximo passo na implementação do algoritmo foi a aplicação do One-Hot Encoding, a qual é uma técnica de pré-processamento amplamente utilizada em machine learning para transformar variáveis categóricas em um formato que pode ser interpretado por algoritmos de aprendizado de máquina. No código em Python essa técnica foi aplicada por meio da ferramenta OneHotEncoder, que pertence a biblioteca scikit-learn, sendo realizado a transformação nas colunas de dados de Education, Occupation, Settlement size, por se tratarem de variáveis categóricas com múltiplas categorias. Após a aplicação dessa técnica, as categorias são transformadas em colunas binárias, onde cada coluna indica a presença ou ausência de uma determinada categoria. Isso ajusta os dados para um formato adequado à aplicação do algoritmo de inteligência artificial.

Em seguida, na implementação do algoritmo foi realizada a normalização das variáveis, uma abordagem de pré-processamento bastante utilizada em Machine Learning para ajustar os dados em uma escala específica. No código em Python implementado, essa técnica foi aplicada por meio da função MinMaxScaler, pertencente à biblioteca scikit-learn, sendo realizada a transformação nas colunas 'Income' e 'Age', que contêm variáveis numéricas contínuas. Após a aplicação dessa técnica, a base de dados resultante apresenta as colunas normalizadas, tornando os dados mais adequados ao processo de aplicação do algoritmo de inteligência artificial.

No desenvolvimento do algoritmo de utilizando o método de K-means, é necessário encontrar o número de clusters(k) ideal, para isso foi implementado no código em Python que realiza criação de uma lista chamada wcss (Within-Cluster Sum of Squares), que armazena os valores de inércia para diferentes números de clusters, variando de 1 a 10. A inércia (ou WCSS) mede a soma das distâncias quadradas entre cada ponto e o centroide do cluster. Na Figura 8 podemos ver a o gráfico plotado utilizando este método onde o eixo x é o número de clusters e o eixo y os valores de WCSS. No gráfico é possível visualizar o ponto onde há uma mudança significativa na redução de WCSS, onde vemos um "cotovelo", esse

ponto é considerado o valor ideal de k, pois representa um equilíbrio entre a complexidade do modelo e a eficiência do agrupamento, sendo essencial a utilização do valor ideal de k no algoritmo a ser implementado para otimizar o desempenho do K-means.

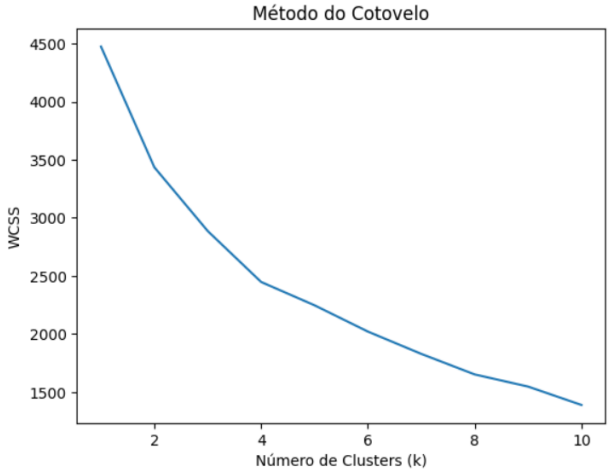


Figura 10 - Fonte: Próprios Autores

O próximo passo na implementação do algoritmo foi a aplicação do PCA (Principal Component Analysis), um método clássico para reduzir a dimensionalidade dos dados, preservando o máximo possível da variância original. No código em Python, essa técnica foi aplicada usando a biblioteca scikit-learn utilizando a ferramenta PCA. Após a transformação, os dados são reduzidos para um espaço de baixa dimensão, facilitando a análise e a visualização e potencializa a performance de modelos de inteligência artificial utilizando estes dados.

O último passo na implementação do algoritmo foi a aplicação do método K-Means, para a clusterização dos dados. No código em Python o modelo foi implementado fazendo a utilização da biblioteca scikit-learn, onde a mesma possui uma ferramenta específica para implementação deste método chamada KMeans, onde na mesma foi definida o número de clusters que foi encontrado com o método cotovelo, sendo este valor igual a 4, e em seguida o algoritmo realiza junção dos clusters encontrados com a base de dados inicial para facilitar a implementação de análises detalhadas dos resultados encontrados.

Após a aplicação do método K-Means, com o número de clusters encontrados sendo igual a 4, foi projetado o gráfico de clusters. O gráfico da Figura 9 representa a clusterização de clientes utilizando o K-means, com os clusters sendo projetados em um espaço bidimensional.

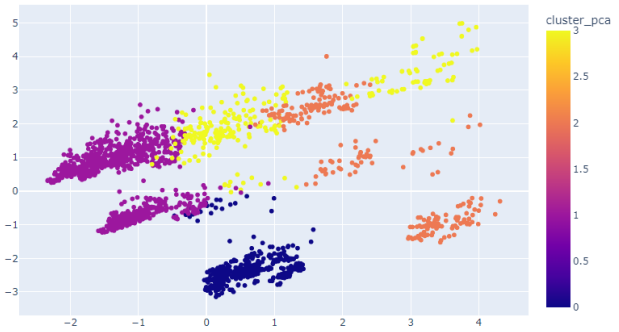


Figura 11 - Gráfico clusterização de clientes

Os pontos no gráfico representam clientes, e cada cor indica o cluster ao qual o cliente pertence. O gráfico mostra 4 clusters distintos, numerados de 0 a 3, de acordo com o número de clusters a partir do Método do Cotovelo. O Cluster 0 (azul escuro), são os clientes mais abaixo, sendo o grupo mais concentrado. O Cluster 1 (roxo) são clientes agrupados à esquerda e abaixo no gráfico, formando um grupo mais coeso. O Cluster 2 (laranja) é um grupo bem definido, ligeiramente separado do restante. Por último, o Cluster 3 (amarelo) é o grupo mais disperso, distribuído em várias regiões do gráfico.

Sobre os padrões visuais, o gráfico mostra que os clusters foram bem separados, o que sugere que o algoritmo K-Means conseguiu agrupar os clientes com base em características que são bastantes distintas. Os Clusters 1 e 0 são bem coesos, indicando que os clientes desses grupos compartilham características bastante próximas. O Cluster 3 apresenta maior dispersão, indicando maior variabilidade entre os clientes neste grupo.

Antes de projetar e evidenciar as características dos clientes de cada cluster, como sexo, estado civil, idade, educação, renda, ocupação e tamanho das cidades, pelo gráfico de clusterização da Figura 9, é possível realizar algumas interpretações.

O Cluster 0 representa um grupo de clientes bastante homogêneo, possivelmente com características semelhantes. O Cluster 1 sugere que esses clientes têm comportamentos ou características muito próximas. O Cluster 2 representa um grupo bem definido e concentrado, cujas características podem diferenciá-los claramente dos outros clusters. Por fim, o Cluster 3 representa o grupo mais disperso, indicando maior diversidade de perfis e comportamentos entre os clientes, o que pode sugerir que esses clientes são menos homogêneos em suas características.

Iniciando a análise dos grupos de clientes com base no cluster, fica evidenciado conforme a Figura 10 e algoritmos utilizando a função value_counts, que o Cluster 1 é o maior grupo, com 974 clientes, o Cluster 0 é composto por 518 clientes, enquanto o Cluster 2 por 256 e o menor grupo é o Cluster 3 com 252 clientes.

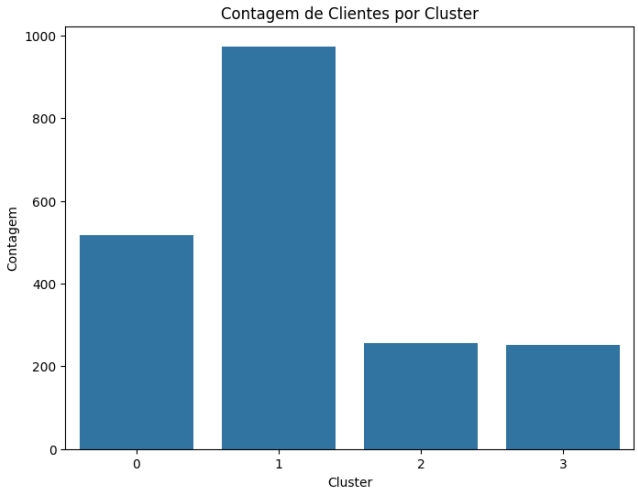
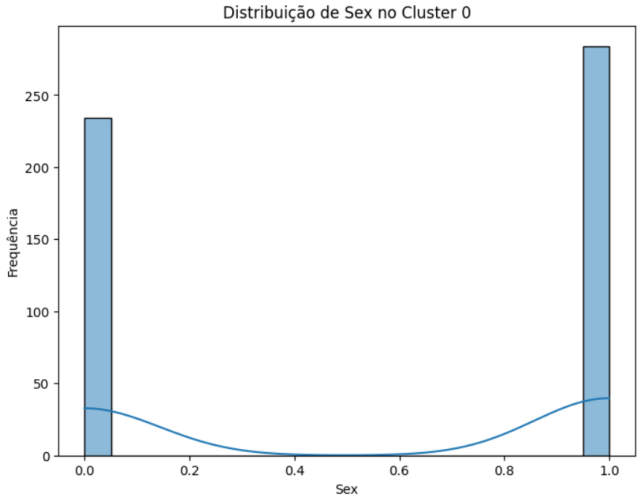
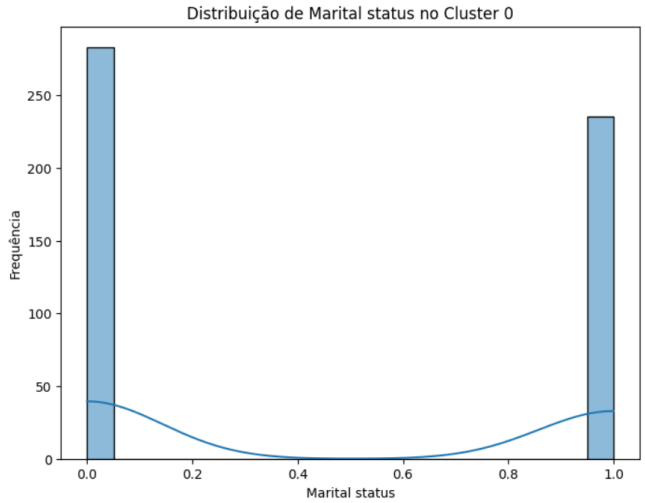
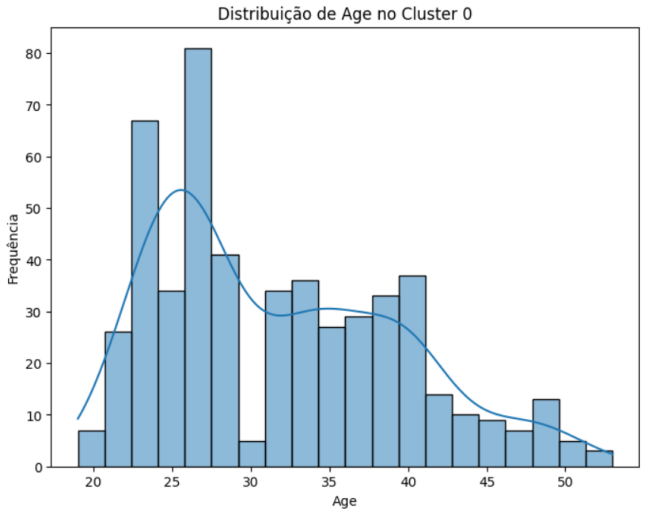


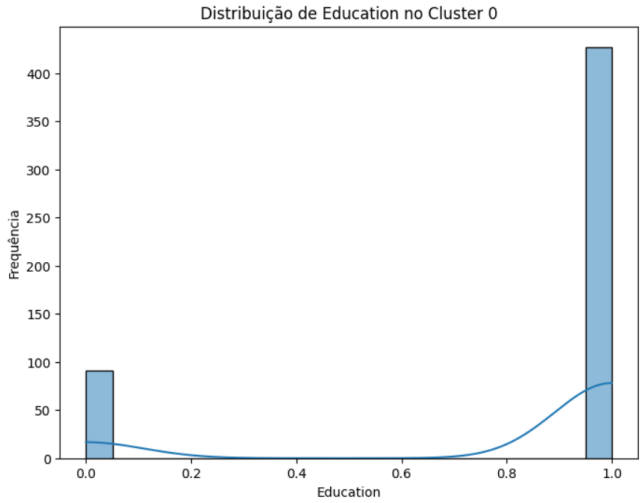
Figura 12 - Gráfico Contagem de Clientes por Cluster

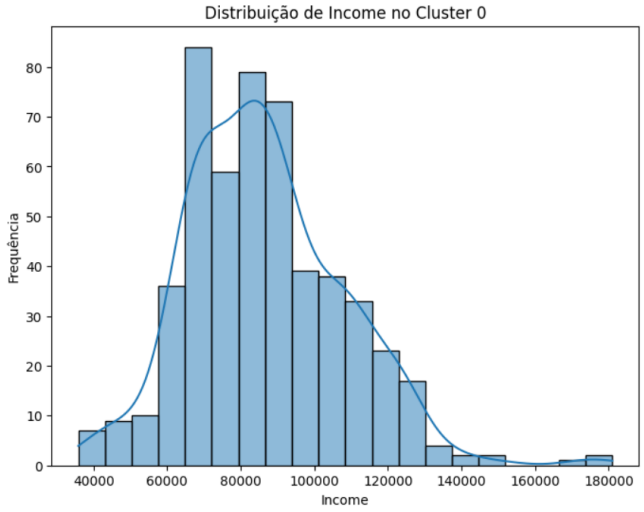
Contudo, para uma abordagem aprofundada, foi realizada uma análise dos grupos de clientes com base nos clusters utilizando proporção para comprovar estatisticamente as análises. Na Figura 11, é evidenciado graficamente as características dos clientes do Cluster 0. Sendo assim, é constatado que os clientes em sua maioria são mulheres (54.83%), solteiros (54.63%), com escolaridade até o high school (82.43%), renda média de $86.700,05, 100% desempregados e vivem em cidades pequenas (95.56%).











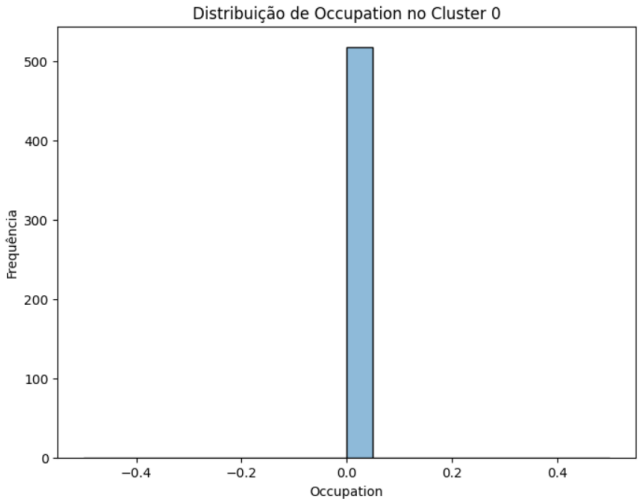
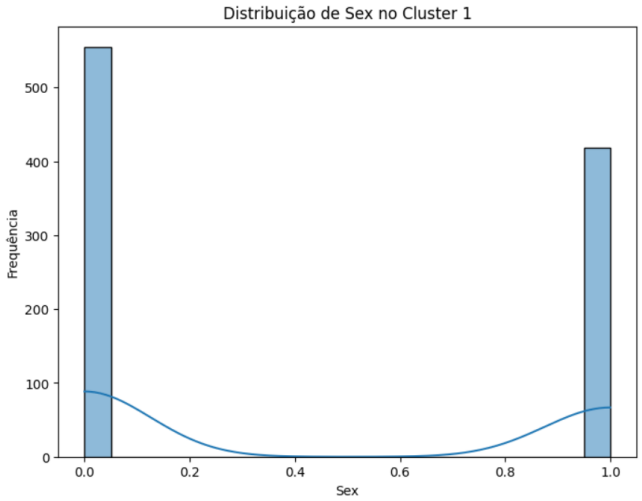
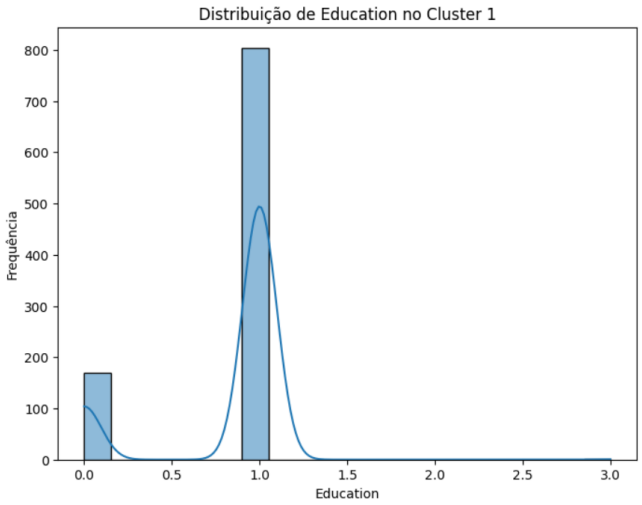
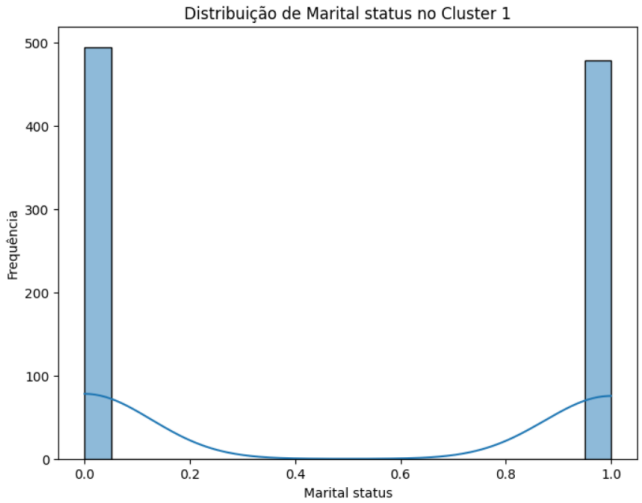


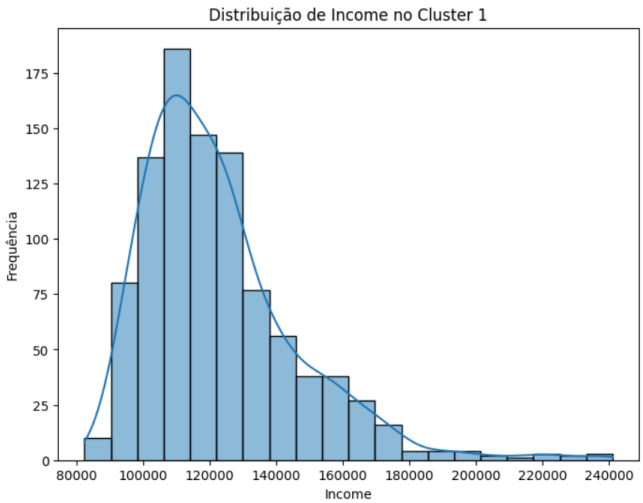
Figura 11 - Gráficos com as características dos clientes do Cluster 0

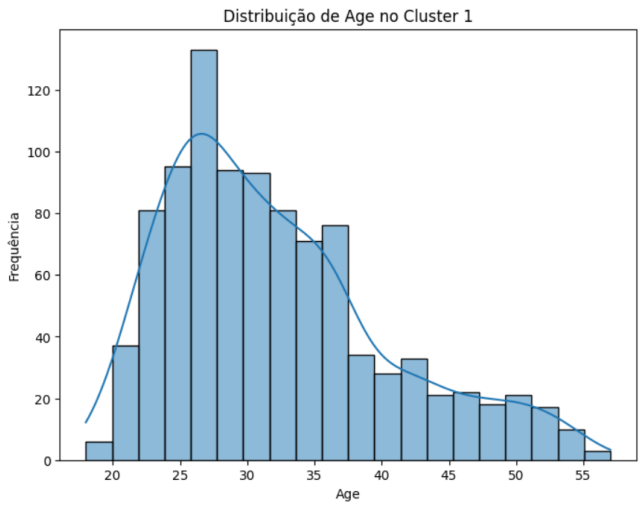
Na Figura 12, é evidenciado graficamente as características dos clientes do Cluster 1. Desta forma, é visto que os clientes em sua maioria são homens (56.98%), com estado civil bem dividido, sendo 50.82% solteiros e 49.18% não solteiros, com escolaridade até o high school (82.55%), renda média de $122.818,20, quase totalmente skilled employee/official (99.08%) e são bem distribuídos em relação a tamanho da cidade, sendo 36.86% os que vivem em cidades de tamanho médio, 36.14% em cidades pequenas e 27.00% em cidades grandes

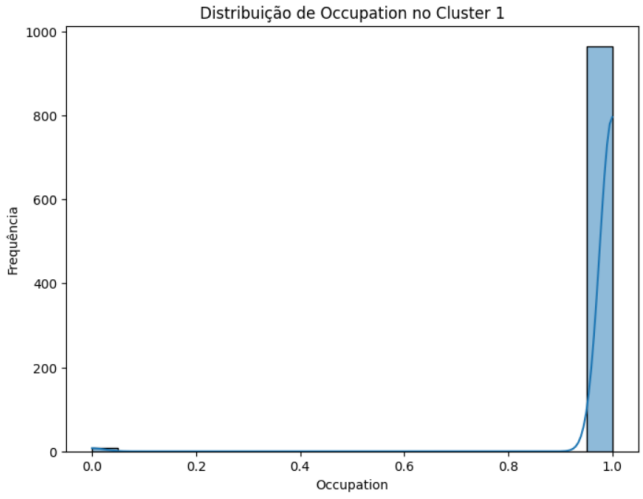

Distribuição de Sex no Cluster 1
Frequência
Sex

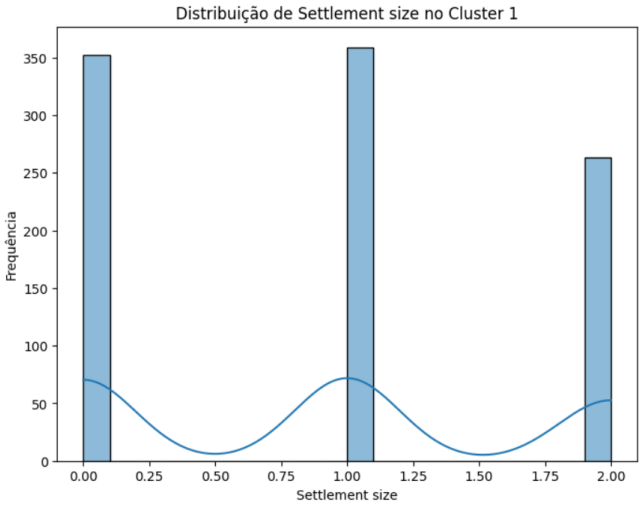

Distribuição de Education no Cluster 1
Frequência
Education


Distribuição de Marital status no Cluster 1
Frequência
Marital status


Distribuição de Income no Cluster 1
Frequência
Income


Distribuição de Age no Cluster 1
Frequência
Age


Distribuição de Occupation no Cluster 1
Frequência
Occupation



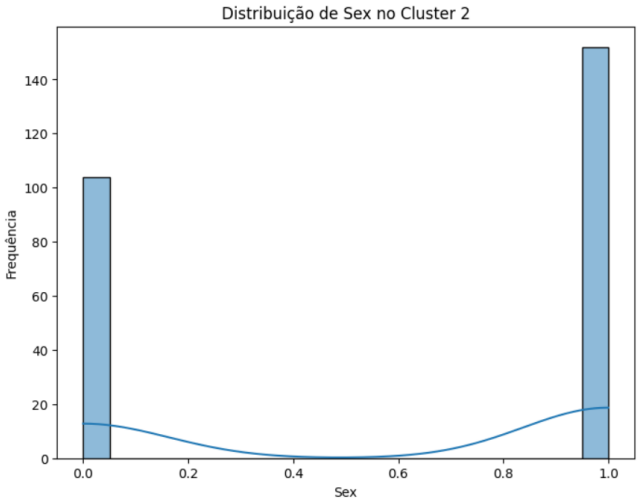
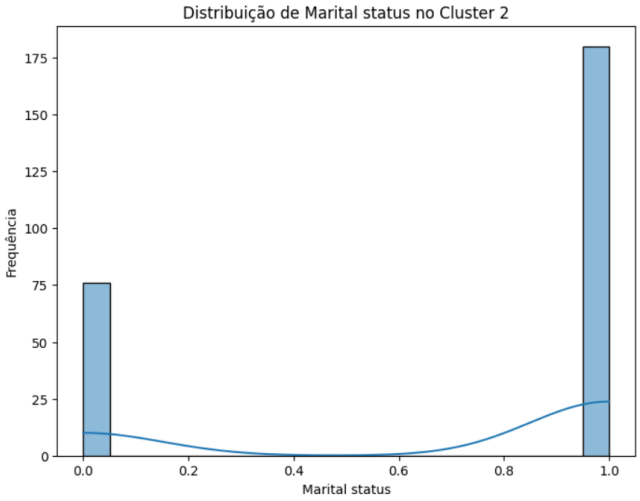
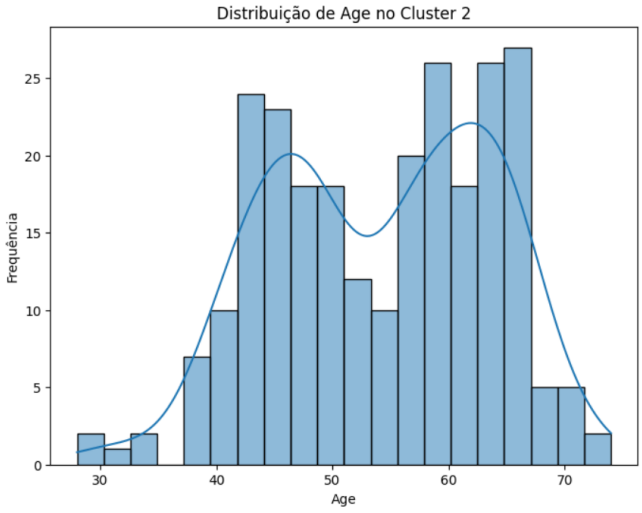
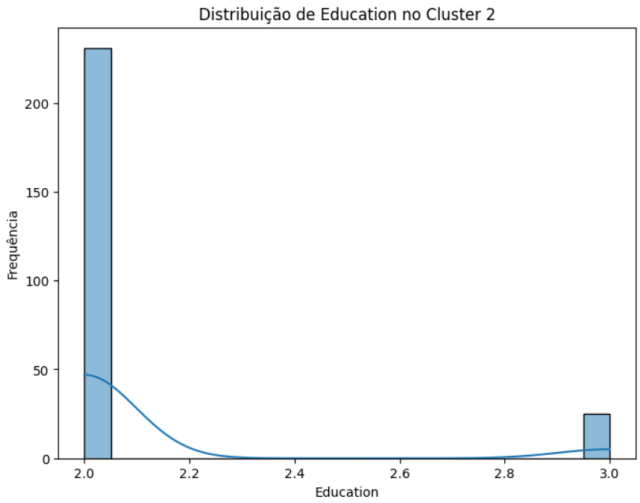


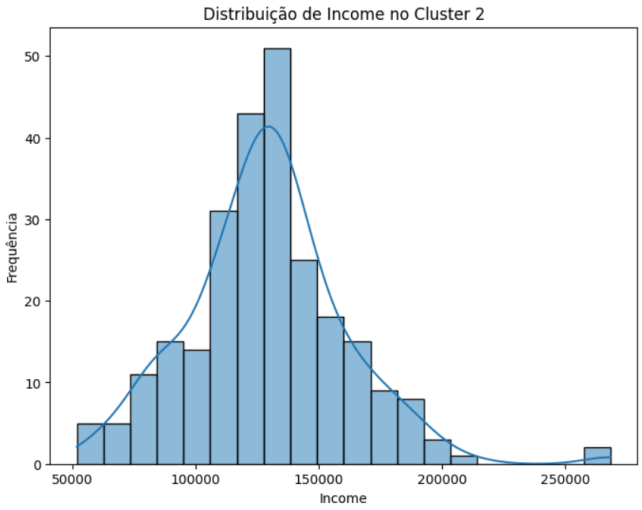
Figura 12 - Gráficos com as características dos clientes do Cluster 1

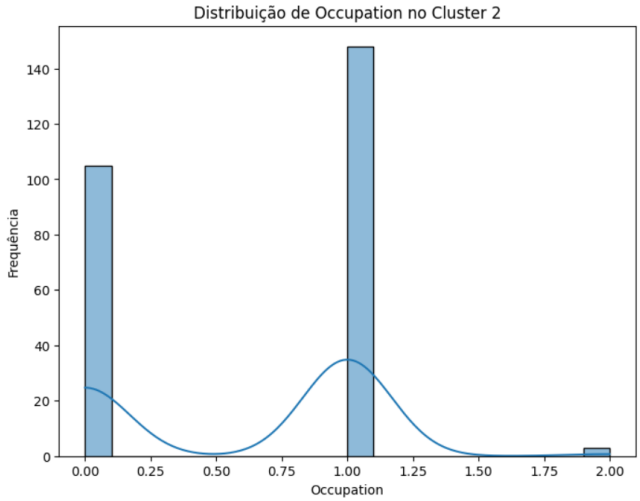
Na Figura 13, é evidenciado graficamente as características dos clientes do Cluster 2. Assim sendo, é observado que os clientes em sua maioria são mulheres (59.38%), não-solteiros (70.31%), com graduação feita em university (90.23%), renda média de $129.024,33, skilled employee/official (57.81%) e vivem em cidades pequenas (50.78%).











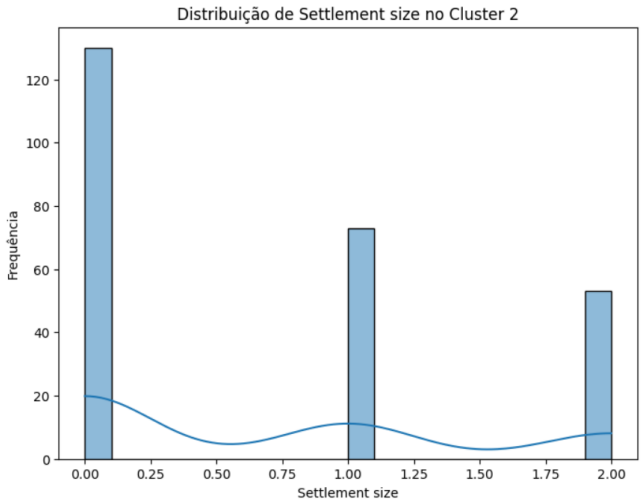
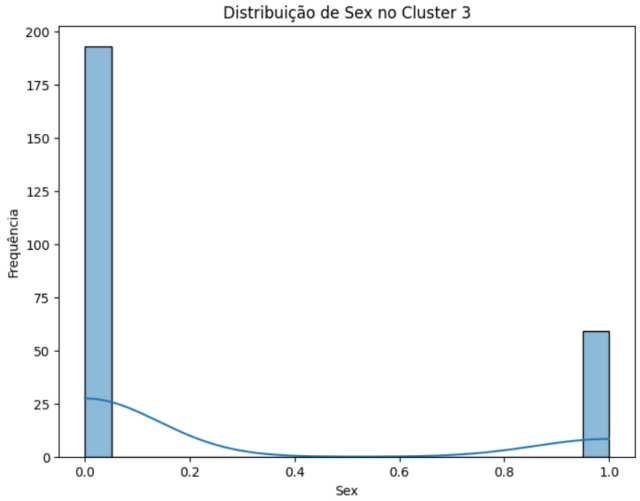
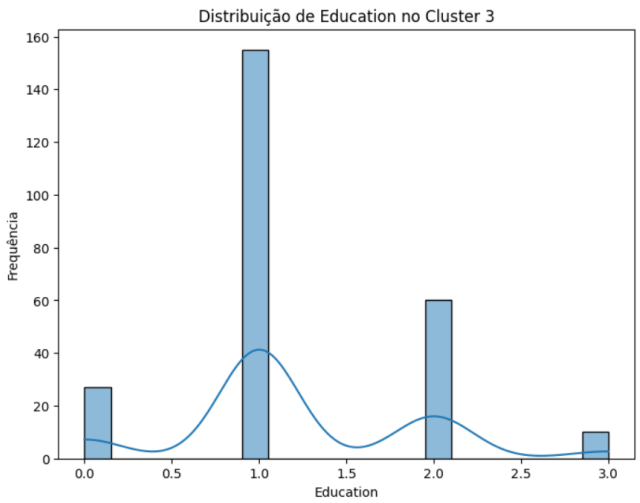
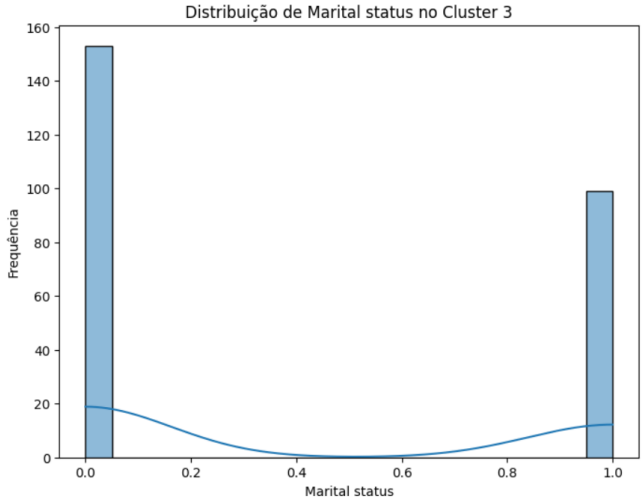


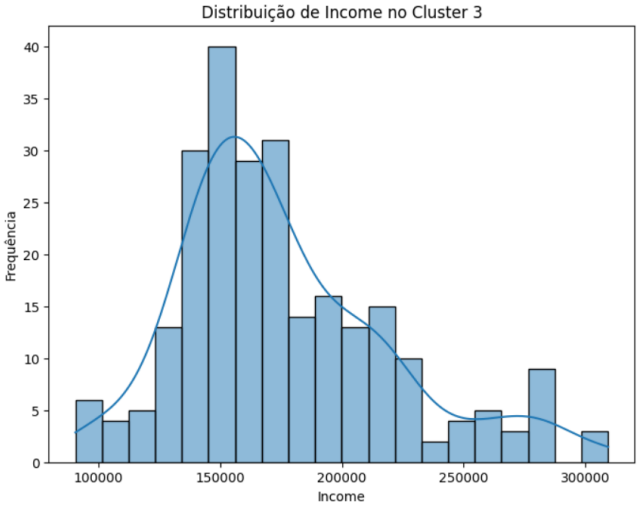
Figura 13 - Gráficos com as características dos clientes do Cluster 2

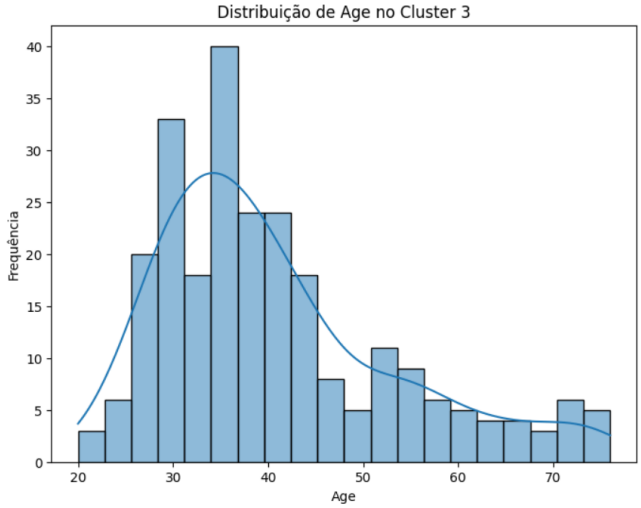
Na Figura 14, é exposto graficamente as características dos clientes do último grupo, do Cluster 3. Deste modo, é verificado que os clientes em sua maioria são homens (76.59%), solteiros (60.71%), com escolaridade até o high school (61.51%), renda média de $175.964,51, com ocupação em management / self-employed / highly qualified employee / officer quase em sua totalidade (99.60%) e vivem em cidades grandes (55.56%).

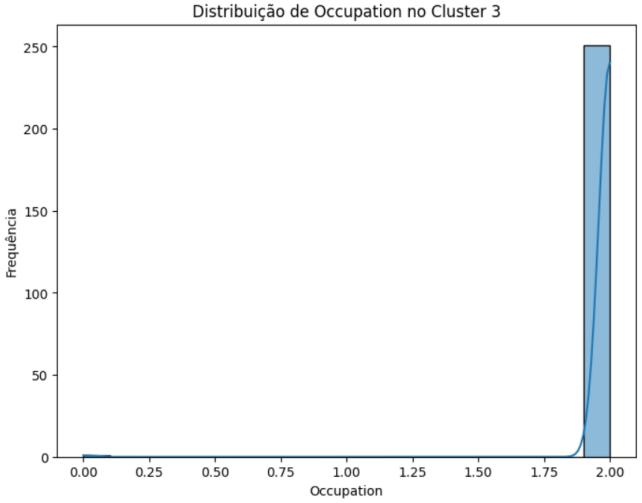

Distribuição de Sex no Cluster 3
Frequência
Sex


Distribuição de Education no Cluster 3
Frequência
Education


Distribuição de Marital status no Cluster 3
Frequência
Marital status


Distribuição de Income no Cluster 3
Frequência
Income


Distribuição de Age no Cluster 3
Frequência
Age


Distribuição de Occupation no Cluster 3
Frequência
Occupation

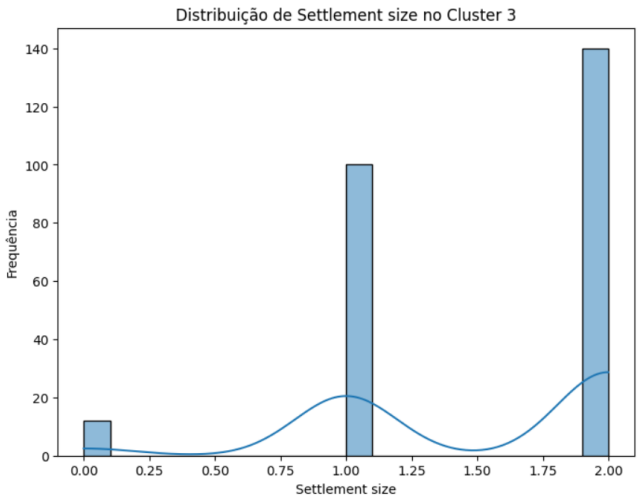


Figura 14 - Gráficos com as características dos clientes do Cluster 3

Desta forma, algumas soluções são possíveis para cada grupo de clientes, visando uma estratégia para a loja:

CLUSTER 0: Focar em produtos acessíveis e oferecer entregas com condições especiais ou estratégias como frete grátis para atrair esses clientes, considerando o possível desafio logístico nas áreas menores.

CLUSTER 1: Ajustar suas ofertas e campanhas com base na diversidade de tamanhos das cidades, oferecendo estratégias logísticas eficazes para as áreas menos urbanizadas.

CLUSTER 2: Adaptar suas campanhas e ofertas para atender às necessidades desse público em áreas menos urbanas, oferecendo produtos que mesclem qualidade e acessibilidade. O marketing pode focar em facilidade de compra online e entrega eficiente.

CLUSTER 3: Para esse grupo de clientes, produtos de luxo e exclusividade são os mais indicados, com foco em serviços premium, como entrega expressa e personalização de produtos.

Os Clusters 0 e 2, sendo compostos por pessoas em cidades pequenas, exigem uma abordagem diferente, voltada para logística eficiente e produtos acessíveis, enquanto os Clusters 1 e 3 podem receber estratégias diferenciadas de acordo com a renda e o tamanho das cidades em que estão concentrados.

5. CONCLUSÃO

Este estudo teve como objetivo aplicar o algoritmo de clusterização K-Means para segmentar clientes utilizando a base de dados Customer Clustering do Kangle. A análise permitiu identificar perfis de clientes distintos com base em características demográficas, financeiras e ocupacionais, o que proporciona à empresa uma oportunidade valiosa de personalizar suas ofertas e estratégias de marketing. Os resultados obtidos mostraram que a clusterização é uma ferramenta eficiente para explorar padrões nos dados e promover uma segmentação mais precisa, possibilitando a criação de estratégias direcionadas e aumentando a probabilidade de sucesso nas interações comerciais.

O uso do K-Means demonstrou-se eficiente na definição de grupos com características similares, embora tenha evidenciado alguns desafios conhecidos, como a dependência da escolha inicial dos centróides e a necessidade de definição prévia do número de clusters. Essa dependência pode gerar resultados diferentes dependendo do ponto de partida, evidenciando a necessidade de utilizar critérios como o Método do Cotovelo para garantir a escolha adequada do número de clusters.

Uma contribuição importante deste trabalho está na capacidade de utilizar a segmentação como um meio de melhorar a experiência do cliente, personalizando ofertas de acordo com suas necessidades e comportamentos específicos. Ao identificar grupos de clientes com maior propensão a gastar ou com maior potencial de fidelização, a empresa pode otimizar seus recursos e direcionar esforços para manter clientes valiosos, ao mesmo tempo em que investe em estratégias para reter aqueles que podem estar em risco de churn. Este tipo de segmentação orientada por dados fornece uma vantagem competitiva significativa, especialmente em mercados cada vez mais dinâmicos e competitivos.

É fundamental explorar o uso de algoritmos que combinem a robustez do K-Means com outras técnicas de aprendizado de máquina, que pode ser mais eficaz em lidar com outliers. Além disso, a implementação de técnicas de redução de dimensionalidade, como a Análise de Componentes Principais (PCA), pode ajudar a visualizar melhor os dados de alta dimensionalidade e identificar características subjacentes que podem não ser evidentes em uma análise mais superficial. Com esses aprimoramentos, os resultados da clusterização são ainda mais precisos e úteis para a tomada de decisões estratégicas.

6. REFERENCES

[1] BAILEY, Christine et al. Segmentation and customer insight in contemporary services marketing practice: why grouping customers is no longer enough. Journal of Marketing Management, v. 25, n. 3-4, p. 227-252, 2009.

[2] CHEN, H.; JUNG, M.; WANG, Y. Predictive modeling in customer churn management: applications in telecom services. Journal of Service Research, v. 22, n. 3, p. 265-278, 2019. [Pode ser omitido se não for mais relevante para o contexto].

[3] GARCÍA, S.; LUENGO, J.; HERRERA, F. Data Preprocessing in Data Mining. Intelligent Systems Reference Library, 2015.

[4] GU, P.; CHENG, F.; YE, X. Enhancing K-Means clustering algorithm with improved initial centroids selection. International Journal of Engineering and Technology, v. 5, n. 4, p. 410-415, 2013.

[5] JOLLIFFE, I. T., & CADIMA, J. Principal Component Analysis: A Review and Recent Developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065), 20150202. DOI: 10.1098/rsta.2015.0202.

[6] KASHWAN, Kishana R.; VELU, C. M. Customer segmentation using clustering and data mining techniques. International Journal of Computer Theory and Engineering, v. 5, n. 6, p. 856, 2013.

[7] KODINARIYA, T. M., & MAKWANA, P. R. Review on determining number of Cluster in K-Means Clustering.

International Journal of Advance Research in Computer Science and Management Studies, 1(6), 90-95.

[8] KETCHEN, David J.; SHOOK, Christopher L. The application of cluster analysis in strategic management research: an analysis and critique. Strategic Management Journal, v. 17, n. 6, p. 441-458, 1996.

[9] KUMAR, S.; AGRAWAL, A.; JAIN, M. A comprehensive study on K-Means clustering algorithm: limitations and solutions. Journal of Artificial Intelligence Research, v. 68, p. 485-512, 2022.

[10] KUAN, Huei-Huang; BOCK, Gee-Woo; VATHANOPHAS, Vichita. Comparing the effects of website quality on customer initial purchase and continued purchase at e-commerce websites. Behaviour & Information Technology, v. 27, n. 1, p. 3-16, 2008.

[11] LEE, Howard B.; MACQUEEN, James B. A K-Means cluster analysis computer program with cross-tabulations and next-nearest-neighbor analysis. Educational and Psychological Measurement, v. 40, n. 1, p. 133-138, 1980.

[12] LEE, J.; KIM, H.; YOON, D. Optimizing customer segmentation through advanced data analytics in retail. Journal of Business Analytics, v. 8, n. 2, p. 123-139, 2020.

[13] LI, Youguo; WU, Haiyan. A clustering method based on K-means algorithm. Physics Procedia, v. 25, p. 1104-1109, 2012.

[14] LUZ, DO da et al. Aplicação de técnicas e mineração de dados e machine learning com Python para análises preditivas sobre o turnover de funcionários: estudo de caso numa empresa multinacional de óleo e gás. SEGET 2023. Disponível em: https://www.aedb.br/seget/arquivos/artigos23/273424.pdf. Acesso em: 18 set. 2024.

[15] MACQUEEN, J. Some methods for classification and analysis of multivariate observations. In: Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley: University of California Press, 1967.

[16] PATRO, S. G. K., & SAHU, K. K. Normalization: A Preprocessing Stage. arXiv preprint arXiv:1503.06462. Disponível em: https://arxiv.org/abs/1503.06462.

[17] PETERSEN, R.; SILVA, M.; CARVALHO, T. Marketing digital: uso de cookies e personalização de conteúdo para aumento de conversão. Marketing Science Review, v. 12, n. 1, p. 76-89, 2020.

[18] RAMIREZ, A.; GARCIA, L.; CRUZ, F. Behavioral segmentation and customer loyalty: the impact of targeted marketing campaigns. Journal of Retailing and Consumer Services, v. 30, p. 95-104, 2021.

[19] REDDY, Biyyapu Sri Vardhan et al. Customer segmentation analysis using clustering algorithms. In: International Conference on Machine Learning, IoT and Big Data. Singapore: Springer Nature Singapore, 2023. p. 353-368.

[20] REDDY, P.; SHARMA, D.; GUPTA, R. Advancements in customer segmentation through K-Means and hybrid clustering techniques. Journal of Data Science and Business Intelligence, v. 15, n. 2, p. 353-368, 2023.

[21] RANA, A.; MALIK, S. Personalization in e-commerce: an empirical study on the impact of behavior-based recommendations. International Journal of E-Commerce Studies, v. 19, n. 4, p. 443-458, 2021.

[22] SHARMA, N.; REDDY, P. Hybrid K-Means clustering: an improved approach to customer segmentation. Journal of Advanced Computational Research, v. 12, n. 3, p. 243-256, 2020.

[23] SINGH, A.; PATEL, J.; KUMAR, R. Machine learning applications in customer segmentation: trends, techniques, and challenges. Expert Systems with Applications, v. 170, p. 114-135, 2021.

[24] WANG, T.; ZHANG, X. Customer profiling using clustering algorithms: improving the accuracy of targeted marketing. Journal of Marketing Analytics, v. 8, p. 97-111, 2020.

[25] ZHANG, L.; LI, Q.; FENG, H. Improving website conversion rates with personalized cookies and behavioral tracking. International Journal of Marketing Technology, v. 11, n. 2, p. 214-227, 2019.

[26] ZHOU, Y.; WANG, J.; ZHENG, T. Principal component analysis in customer segmentation: reducing dimensionality for effective cluster analysis. Journal of Data Mining and Business Applications, v. 7, n. 1, p. 45-61, 2021.

[27] ZHU, Y.; CHEN, F.; LIU, J. E-commerce personalization through customer behavior analysis: a case study of Amazon.com. Journal of E-Business Research, v. 13, p. 73-85, 2020.

[28] ZHENG, A.; CASARI, A. Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists. O'Reilly Media, 2018.