**UNIVERSIDADE PAULISTA**

**ADRIANO SILVA GAMA**

**ANDERSON VIEIRA FARAGO**

**DIEGO DE LIMA HONDA**

**ERICK OLIVEIRA KAWAUCHE**

**MAURO SANCHES FREITAS**

**ANÁLISE DE AGRUPAMENTO DE PERFIS DE CLIENTES DE UM COMÉRCIO ATACADISTA UTILIZANDO O ALGORITMO K-MEANS**

**Alphaville**

**2017**

**ADRIANO SILVA GAMA**

**ANDERSON VIEIRA FARAGO**

**DIEGO DE LIMA HONDA**

**ERICK OLIVEIRA KAWAUCHE**

**MAURO SANCHES FREITAS**

**ANÁLISE DE AGRUPAMENTO DE PERFIS DE CLIENTES DE UM COMÉRCIO ATACADISTA UTILIZANDO O ALGORITMO K-MEANS**

Trabalho de conclusão de curso para obtenção do título de graduação em Sistema de Informação apresentado à Universidade Paulista – UNIP.

Orientador: Prof.ª Ana Carolina

**Alphaville**

**2017**

CIP - Catalogação na Publicação

ANÁLISE DE AGRUPAMENTO DE PERFIS DE CLIENTES DE UM COMÉRCIO ATACADISTA UTILIZANDO O ALGORITMO

K-MEANS / Erick Oliveira Kawauche...[et al.]. - 2017.

45 f. : il. color

Trabalho de Conclusão de Curso (Graduação) apresentado ao Instituto de Ciência Exatas e Tecnologia da Universidade Paulista, Alphaville, 2017.

Área de Concentração: Inteligência Artificial. Orientador: Prof. LD Ana Carolina Bueno Borges.

1. Aprendizado de máquina. 2. K-means. 3. Agrupamento. 4. Segmentação de clientes. I. Kawauche, Erick Oliveira. II. Borges, Ana Carolina Bueno (orientador).

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da Universidade Paulista com os dados fornecidos pelo(a) autor(a).

**ADRIANO SILVA GAMA**

**ANDERSON VIEIRA FARAGO**

**DIEGO DE LIMA HONDA**

**ERICK OLIVEIRA KAWAUCHE**

**MAURO SANCHES FREITAS**

**ANÁLISE DE AGRUPAMENTO DE PERFIS DE CLIENTES DE UM COMÉRCIO ATACADISTA UTILIZANDO O ALGORITMO K-MEANS**

Trabalho de conclusão de curso para obtenção do título de graduação em Sistema de Informação apresentado à Universidade Paulista – UNIP.

Aprovado em:

BANCA EXAMINADORA

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_/\_\_/\_\_\_

Prof.

Universidade Paulista – UNIP

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_/\_\_/\_\_\_

Prof.

Universidade Paulista – UNIP

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_/\_\_/\_\_\_

Prof.

Universidade Paulista UNIP

**AGRADECIMENTOS**

Primeiramente а Deus que permitiu que tudo isso acontecesse ao longo de minha vida;

À minha família que acreditou em meu potencial e me deu total suporte durante toda a minha jornada;

Meus sinceros agradecimentos a minha grande amiga Cristine Guadelupe, que me apoiou não somente na elaboração deste trabalho, mas também me incentivou nas horas difíceis de desânimo e cansaço;

**Resumo**

Um dos grandes desafios de qualquer empresa hoje em dia é entender as necessidades dos seus clientes, e com base nesse conhecimento conseguir oferecer produtos que realmente são interessantes para cada indivíduo. A compreensão do comportamento de cada cliente de forma individual se tornou uma peça fundamental para maximizar o lucro. As empresas que assumem esse desafio estão aprendendo a segmentar seus clientes e ter uma melhor visão do seu público alvo, sendo assim, elas tem uma noção maior nos clientes em que devem investir e naqueles que não valem a pena o investimento. Para encontrar os padrões comportamentais dos clientes são utilizadas técnicas como aprendizagem de máquina e data mining e o resultado dessa junção permite que uma empresa melhore a forma como seu marketing é feito, como as compras são conduzidas e qual a melhor forma de abordar o cliente de forma a melhor atender suas necessidades. Este trabalho apresenta uma análise detalhada de uma base de dados dos clientes de um atacado, o objetivo deste projeto é, através das técnicas de aprendizado de máquina segmentar os clientes do atacado em diferentes perfis, que dará ao distribuidor discernimento sobre como melhor estruturar seu serviço para atender as necessidades de cada cliente. A base de dados passou por um pré-processamento e limpeza, que incluem escalonamento, utilizando o logaritmo natural, detecção de pontos discrepantes e utilização da técnica de análise de componentes principais para reduzir a dimensão da base de dados. Após o pré-processamento os dados foram submetidos ao algoritmo de agrupamento conhecido como K-Means, este algoritmo é responsável por realizar a segmentação com base nas características similares entre cada dado.

Palavras-chaves: Aprendizado de máquina. K-Means. Agrupamento. Segmentação de clientes.

**Abstract**

One of the biggest challanges of every forward-looking companies it's to understand the needs of each client, and using that understanding to offer products that are in the best interest of each customer. The understanding of the behavior of each customer is the key to maximize the proft. These same companies are learning to look at the values of each customer, so they know which ones are worth investing money and effort to hold on to, and which ones that are not profitable. In order to find patterns behaviors of a set of customers we use techniques like machine learning and data mining, the results allow a corporation to improve its marketing, sales and customer support. This project present a detailed analysis of a wholesale customer data set, and its objective it's to use machine learning to cluster the customers of the wholesale in different groups, which will provide knowledge to the wholesale distributor so it can have a better understanding of the customers needs, and improve it sales. The data set used in this project have gone through a variaty of preprocessing methods, that includes feature scaling, identifying outliers and removing them and finally applying the Principal Components Analysis in order to make it ready to apply the K-Means algorithm, which is responsable to cluster a data set into distinct types of groups based on each observation point similarities. The result of this project is presented as a well clustered set of distinct types of customers that are present in the data set. Each cluster presented by the K-Means algorithm makes it clear the differences between each customer. As a result we conclude that the K-Means algorithm brought a positive solution to our client’s clusterization problem. The cluster presented by the algorithm was well defined as well as the distinct type of clients presente in our initial dataset.

Keywords: Machine learning, K-Means, Cluester. Client Segmentation.

[1 Introdução 9](#_Toc500189390)

[1.1 Problema 10](#_Toc500189391)

[1.2 Objetivo 10](#_Toc500189392)

[1.3 Hipótese 11](#_Toc500189393)

[1.4 Justificativa 11](#_Toc500189394)

[2 Fundamentos Teóricos 13](#_Toc500189395)

[2.1 Conhecimento através da mineração de dados 13](#_Toc500189396)

[2.1.1 Seleção dos Dados 14](#_Toc500189397)

[2.1.2 Pré-processamento e limpeza 14](#_Toc500189398)

[2.1.3 Transformação dos dados 16](#_Toc500189399)

[2.1.4 Data Mining 16](#_Toc500189400)

[2.2 Aprendizado de Máquina 17](#_Toc500189401)

[2.2.1 Aprendizado Não Supervisionado 17](#_Toc500189402)

[2.2.2 Redução de Dimensionalidade 19](#_Toc500189403)

[2.2.3 Algoritmo de Agrupamento (Clustering) 23](#_Toc500189404)

[2.2.4 Algoritmo de Agrupamento: K-Means 23](#_Toc500189405)

[2.2.5 Funcionamento do algoritmo K-means 24](#_Toc500189406)

[3 Desenvolvimento 26](#_Toc500189407)

[3.1 Python 26](#_Toc500189408)

[3.2 Base de dados 26](#_Toc500189409)

[3.3 Observação dos Dados 27](#_Toc500189410)

[3.3.1 Análise dos Dados 27](#_Toc500189411)

[3.3.2 Resultado da análise dos dados 30](#_Toc500189412)

[3.4 Pré-Processamento dos dados 31](#_Toc500189413)

[3.4.1 Escalonamento dos dados 31](#_Toc500189414)

[3.4.2 Detecção de Desvios 33](#_Toc500189415)

[3.5 Transformação dos Atributos 34](#_Toc500189416)

[3.5.1 Análise das dimensões do PCA 35](#_Toc500189417)

[3.5.2 Conclusão da análise das dimensões 36](#_Toc500189418)

[3.6 Algoritmo de agrupamento (K-Means) 37](#_Toc500189419)

[3.6.1 Método do cotovelo (Elbow Method) 37](#_Toc500189420)

[3.6.2 Método do coeficiente de silhueta 39](#_Toc500189421)

[3.6.3 K-Means 40](#_Toc500189422)

[3.7 Data Mining 41](#_Toc500189423)

[4 Conclusão 42](#_Toc500189424)

[5 Bibliografia 43](#_Toc500189425)

[APÊNCIDE A – Código fonte do desenvolvimento do agrupamento de perfis dos clientes 45](#_Toc500189426)

# Introdução

O número de supermercados e lojas de conveniências crescem cada vez mais, porém o número de clientes não estão aumentando da mesma maneira (KHANAL, 2016, p. 2). Khanal (2016), observa que antigamente o preço era o único diferencial entre os supermercados, já hoje em dia muitos outros fatores são levados em consideração. A utilização de novas tecnologias é um fator importante para aprimorar o modelo de negócio e a satisfação do cliente.

Atualmente, todos os setores comerciais estão focando em entender a necessidade de cada cliente individualmente, tornando mais fácil atendê-los, e não se desgastando com os competidores (BERRY e LINOFF, 2004, p. 2).

Com essa mudança de mentalidade de focar mais no cliente, essas empresas estão aprendendo a avaliar quais clientes possuem mais valor para ser investido (BERRY e LINOFF, 2004, p. 2). Através das técnicas de aprendizagem de máquina é possível encontrar relações entre o comportamento do cliente e como isso afeta suas compras. Nos supermercados por exemplo, leite e pão são relacionados entre si, por isso seria melhor colocá-los um perto do outro na prateleira (KHANAL, 2016, p. 4).

Hoje em dia tem se tornado muito comum recebermos anúncios direcionados com base em compras passadas. Isso acontece não só em sites na internet, os supermercados também utilizam dados dos clientes para decidir quais produtos oferecer para eles em futuras compras. Berry e Linoff (2004) observam que as informações sobre o comportamento de cada cliente são como os olhos, ouvido, olfato e tato da empresa.

De acordo com Berry e Linoff (2004), os dados passados de um cliente contém muitas informações que podem ser úteis no futuro. Isso acontece pois essas informações sobre compras, preferências e comportamentos dos clientes não são randômicas mas refletem as necessidades dos clientes.

## Problema

Segundo Ohta e Higuchi (2013), uma experiência de compra agradável e satisfatória para o cliente está relativamente ligado a facilidade de como os produtos são encontrados e o tempo de permanência na loja. Ou seja para o cliente, quanto menos tempo ele gasta dentro do mercado e quanto mais fácil for para encontrar os produtos que precisa, é melhor. Em contrapartida para o mercado é mais interessante oferecer novos produtos para promover novos hábitos sem que o cliente sinta desinteressado ou tenha a sensação que está vagando por horas dentro do mercado (OHTA e HIGUCHI, 2013, p. 1).

Levando em consideração o cliente e o vendedor, o layout do mercado se torna um ponto chave para atender ambos os lados. Cil, Ay e Turkan nota que o layout de uma loja é uma chave importante não só para satisfazer as necessidades do cliente mas também para influenciar suas preferências e necessidades.

Diante desse problema seria mais interessante se o estabelecimento conseguisse segmentar seus clientes definindo um perfil de compra para cada um deles. Dessa maneira seria possível atender melhor cada um desses perfis com maior precisão, oferecendo o que realmente interessa e sugerindo novos produtos que se encaixe no perfil de cada um deles.

## Objetivo

O objetivo deste trabalho é analisar um conjunto de dados referentes as despesas de vários clientes em um atacado, e com base nas despesas e nos diferentes produtos comprados segmentar os clientes em diversos perfis. Isso dará ao distribuidor discernimento sobre como melhor estruturar seu serviço para atender as necessidades de cada cliente. Ao fim do desenvolvimento apresentaremos um modelo de aprendizado de máquina capaz de segmentar todos os clientes em suas devidas categorias.

O escopo deste projeto se restringe em segmentar o perfil do cliente considerando suas compras passadas, não será levado em consideração a relação dos produtos comprados ou qual será a disposição desses produtos dentro do estabelecimento. A base dados são referentes a clientes de um atacado, ou seja, possivelmente são revendedores, supermercados, restaurantes ou hotéis, por este motivo não são clientes finais.

Todo o código fonte utilizado para o desenvolvimento deste projeto está anexado ao apêndice A.

## Hipótese

Para solucionar o problema de proporcionar uma melhor experiência de compra para o cliente e ainda assim manter os interesses do mercado, sugerimos a segmentação do estabelecimento baseado nos tipos de compras realizadas pelos clientes. Por exemplo criar uma seção de compras rápidas, para clientes que estão fazendo compras de última hora, ou criar uma seção para itens de animais, para os clientes que estão fazendo compras para seus animais domésticos. A segmentação da loja afeta a aceitabilidade dos preços pelo cliente, os supermercados estão revolucionando o serviço de varejo, em sua essência criando modos mais efetivos para organizar os produtos e incentivar o auto atendimento (CIL, AY e TURKAN, 2009, p. 2).

Para esta solução será utilizado aprendizado de máquina para identificar o perfil de cada cliente, levando em consideração suas compras passadas. Utilizando o método de agregação será possível entender os comportamentos de compras dos clientes e assim auxiliar na tomada de decisão para segmentar a loja, sugerir produtos de interesse do cliente ou até mesmo criar promoções direcionadas para cada perfil de cliente.

## Justificativa

Segundo Berry e Linoff (2004) todas as empresas estão caminhando para um único objetivo, o de entender as necessidades de cada cliente individualmente. E as empresas estão utilizando esse entendimento para fidelizar os clientes ao invés de competir diretamente com os concorrentes.

As empresas que assumem esse objetivo estão aprendendo a segmentar seus clientes e ter uma melhor visão do seu público alvo, sendo assim, elas tem uma noção maior nos clientes em que devem investir e naqueles que não valem a pena o investimento. (BERRY e LINOFF, 2004, p. 2).

Para encontrar os padrões comportamentais dos clientes são utilizadas técnicas como por exemplo aprendizagem de máquina e data mining e de acordo com Berry e Linoff (2004) o resultado dessa junção permite que uma empresa melhore a forma como seu marketing é feito, como as compras são conduzidas e qual a melhor forma de abordar o cliente de forma a melhor atender suas necessidades.

# Fundamentos Teóricos

Nesta seção serão apresentados os conceitos teóricos e estudos que utilizamos como base para alcançar o objetivo deste projeto. Os próximos tópicos apresentam um conceito básico sobre mineração de dados e sua importância nos estudos de aprendizado de máquina.

## Conhecimento através da mineração de dados

De acordo com Fayyad, Piatetsky-Shapiro e Smyth (1996), a velocidade com que as informações estão sendo acumuladas são drásticas, e existe uma grande necessidade de novos métodos para extrair informações relevantes destes grandes volumes de dados. Uma dessas técnicas é chamado de Descoberta de Conhecimento em Banco de Dados (em inglês Knowledge Discovery in Databases, KDD).

KDD é definido como “[...] um processo, não trivial, de extração de informações implícitas, previamente desconhecidas e potencialmente úteis, a partir dos dados armazenados em um banco de dados.” (FAYYAD, PIATETSKY-SHAPIRO e SMYTH, 1996, p. 40).

A informação por si só não agrega nenhum valor se não for interpretada de maneira correta, por esse motivo precisamos transformar essa informação em conhecimento. Segundo Fayyad, Piatetsky-Shapiro e Smyth (1996), o processo de Descoberta de Conhencimento em Banco de Dados (Figura 1) pode ser dividido em alguns passos que serão listados nos próximos tópicos.

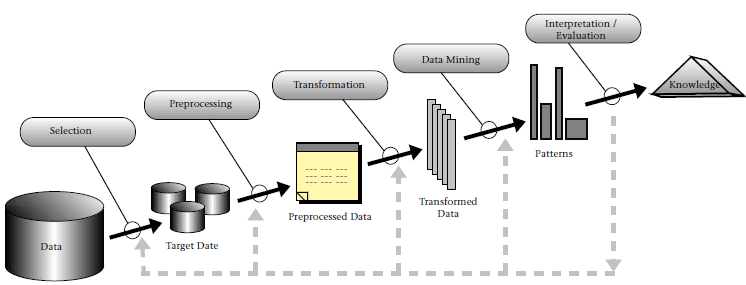


Figura : Resumo do processo que compões o KDD. Fonte: (FAYYAD, PIATETSKY-SHAPIRO e SMYTH, 1996, p. 41)

### Seleção dos Dados

O primeiro passo é entender a aplicação como um todo, e identificar a motivação para realizar o processo de KDD, para isso é necessário entender a visão do cliente e suas necessidades. (FAYYAD, PIATETSKY-SHAPIRO e SMYTH, 1996, p. 42). Com um objetivo em mãos será necessária a seleção dos dados que serão analisados. A seleção tem um impacto significante sobre a qualidade do resultado final, nesta fase é feita a escolha do conjunto de dados com todos os possíveis atributos e registros. O processo de seleção dos dados é bastante complexo, pois os dados podem vir de diferentes repositórios de dados (PRASS, 2017).

Para este trabalho estamos utilizando uma base de dados disponível em um repositório on-line, por este motivo essa etapa de seleção dos dados já foram realizadas.

Depois de realizado a seleção dos dados é necessário iniciar o processo de pré-processamento e limpeza da base de dados, para que depois possa ser aplicado o algoritmo de aprendizado de máquina. Na próxima seção será apresentado como funciona o processo da limpeza e pré-processamento dos dados.

### Pré-processamento e limpeza

A etapa do pré-processamento é um passo muito importante, pois a qualidade dos dados terá influência direta no modelo de aprendizado de máquina. Segundo Prass (2017), o processo de limpeza exige a eliminação de dados redundantes e também o tratamento de dados em brancos ou vazios. Esses ajustes na base de dados evitam que o modelo seja gerado com ruídos que levem a uma interpretação equivocada do resultado final. Durante esse processo é necessário levar em consideração somente os dados que forem considerados relevantes para o estudo, não há a necessidade de processar dezenas de dados se somente alguns serão utilizados. Vale lembrar que os dados que não forem considerados relevante, não necessariamente significam que são dados ruins, dependendo da abordagem tomada os dados descartados podem se tornar relevantes.

Para este trabalho utilizamos duas técnicas de pré-processamento para limpar a base de dados, eles serão apresentados nos seguintes tópicos.

#### Escalonamento dos dados

Durante o desenvolvimento deste projeto utilizamos o algoritmo chamado K-means, que será explicado com mais detalhe mais adiante. Este algoritmo utiliza a fórmula da distância euclidiana para calcular a distância entre cada ponto de observação, logo qualquer variável que estiver em uma escala bem maior terá um efeito muito maior no cálculo da distância, por esse motivo é extremamente importante escalonar os dados antes de aplicar qualquer algoritmo que se baseia na distância (JAMES, WITTEN, *et al.*, 2013, p. 165).

Segundo James, Witten, et al. (2013), o escalonamento dos dados é deixar os dados na mesma escala, como em nossa base de dados estamos processando valores monetários, as escalas dos valores são muito diferentes e isso pode gerar resultados indesejados ao aplicar o algoritmo K-means.

Hamilton (2014), nota que utilizar logaritmo natural para descrever relações entre atributos com valores financeiro é muito mais interessante, pois ao aplicar o logaritmo natural os dados ganham mais significado e se tornam mais fáceis de serem interpretados. Por este motivo decidimos escalonar os dados utilizando o logaritmo natural (log), pois com ele é possível reduzir grandezas elevadas para valores menores

#### Detecção de desvios

Um desvio é um ponto muito distante do conjunto de valores previsto pelo modelo de aprendizado de máquina (JAMES, WITTEN, *et al.*, 2013). Na Figura 2 é possível observar um exemplo de um ponto de desvio, também chamado de *outlier.* O ponto claramente está fora do conjunto de dados e pode ser tratado como um outlier.

Segundo James, Witten, et al. (2013), esses pontos fora da curva acontecem por diversos motivos, como por exemplo um erro durante a coleta de dados. A forma como esses pontos de desvios serão tratados variam de acordo com o modelo de dados utilizados. Caso esteja analisando uma base de dados de cartões de créditos, os pontos de desvios poderiam indicar uma possível fraude. Ou caso os pontos de desvios sejam irrelevantes para o modelo eles podem ser removidos (GÉRON, 2017).

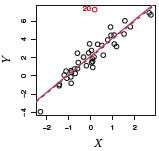


Figura : Exemplo de um desvio em um modelo de regressão linear (JAMES, WITTEN, *et al.*, 2013, p. 97)

### Transformação dos dados

Segundo Prass (2017), Após os dados serem selecionados, limpos e pré-processados eles necessitam ser armazenados em um formato válido para ser aplicado os algoritmos de aprendizagem de máquina, ou seja deve ser transformado para que o algoritmo possa consumir e gerar o resultado final. Nessa fase do processo do KDD é comum encontrar novos conjuntos de dados a partir dos dados existentes, como por exemplo, dado o ano de nascimento do indivíduo é possível saber a idade atual do mesmo.

Durante a transformação dos dados é necessário identificar os dados que são considerados importantes para o objetivo que desejamos alcançar (FAYYAD, PIATETSKY-SHAPIRO e SMYTH, 1996).

### Data Mining

Essa é a etapa final e pode ser melhor definida como “[...] Data Mining é a exploração e análise, de forma automática ou semiautomática, de grandes bases de dados com objetivo de descobrir padrões e regras. O objetivo do processo de mineração é fornecer as corporações informações que as possibilitem montar melhores estratégias de marketing, vendas, suporte, melhorando assim os seus negócios”. (BERRY e LINOFF, 2004, p. 7).

Segundo Fayyad, Piatetsky-Shapiro e Smyth (1996), um dos principais objetivos da mineração de dados podem se resumir em previsão e descrição. A previsão envolve analisar os dados apresentados na base de dados, para prever algum resultado desconhecido. Já a descrição foca em descobrir padrões comportamentais com base em dados passados. No caso deste projeto onde o foco é descobrir os perfis dos clientes de um comercio atacadista, optamos por utilizar o conceito de descrição para encontrar padrões comportamentais dos clientes.

## Aprendizado de Máquina

Nesta seção serão apresentados os conceitos básicos de aprendizado de máquina, com o foco nas técnicas utilizadas para o desenvolvimento deste projeto.

### Aprendizado Não Supervisionado

A maior parte dos problemas de aprendizado de máquina se encaixam em duas categorias: Aprendizado Supervisionado e Aprendizado Não Supervisionado. No modelo de aprendizado supervisionado, para cada dado observado existe uma resposta associada , ou seja, para cada conjunto de dados existe um resultado esperado. De acordo com James, Witten, Hastie e Tibshirani (2013) o modelo de aprendizado supervisionado procura ajustar os dados observados e suas respectivas respostas esperadas com o objetivo de futuramente prever os resultados de um novo conjunto de dados, este modelo tenta buscar uma relação dos dados e seus resultados .

Já no modelo de aprendizado não supervisionado temos a situação onde, para cada dado observado , temos as informações sobre porém não temos uma resposta associada . Não é possível ajustar um modelo de previsão pois não temos uma resposta para prever. Este cenário é denominado como ‘não supervisionado’ pois nos faltam dados de respostas que possibilita a supervisão dos dados observados (JAMES, WITTEN, *et al.*, 2013, p. 26).

De acordo com James, Witten, Hastie e Tibshirani (2013), uma ferramenta muito útil para realizar uma análise estatística dentro deste cenário onde há falta de dados de respostas , é a análise por agrupamento. O objetivo da análise por agrupamento é para assegurar que em um conjunto de dados , os dados observados se agrupem em categorias distintas. Podemos tomar como exemplo a segmentação de um supermercado, onde levamos em consideração diversas características dos clientes, tal como endereço, ganho mensal, quais seus hábitos durante as compras. E a partir disso segmentar os clientes em grupos diferentes, como aqueles que gastam mais, ou gastam menos (JAMES, WITTEN, *et al.*, 2013, p. 27). Se as informações sobre os padrões de gastos de um cliente já fosse revelado, seria possível utilizar um modelo supervisionado para prever quais clientes tem maior tendência em gastar mais ou menos. Como nesse cenário não temos acesso a essas informações, a solução é tentar agrupar os clientes com base em suas características. Podendo assim identificar diferentes grupos de cliente com base em seus pontos de interesses.

Na temos um exemplo de como funciona um modelo por agrupamento. No gráfico foram plotados 150 pontos de observações com base em duas variáveis e , onde cada ponto de observação corresponde a um grupo distinto, para facilitar o exemplo cada grupo foi designado uma cor e um símbolo diferente. Do lado esquerdo é possível observar que os pontos estão nitidamente agrupados, porém na prática nem sempre é assim. Do lado direito mostra um cenário mais próximo do real, onde os grupos estão todos juntos e praticamente não dá para distinguir a qual grupo cada ponto pertence (JAMES, WITTEN, *et al.*, 2013, p. 27).

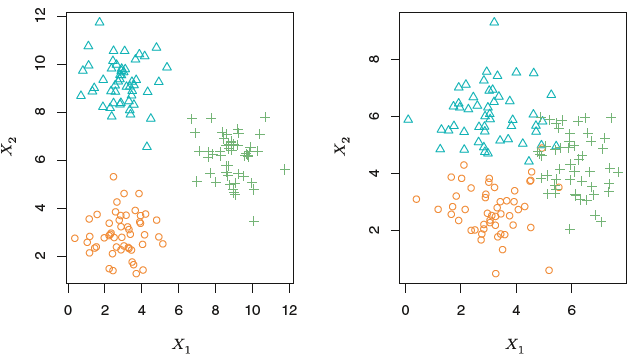


Figura : Exemplo de um modelo de agrupamento. Fonte: (JAMES, WITTEN, *et al.*, 2013, p. 27)

Segundo James, Witten, et al., o modelo não supervisionado é muito mais subjetivo, pois não existe uma diretriz para seguir como referência, isso se deve pelo fato de que não é possível validar se as respostas obtidas no modelo não supervisionado são 100% verdadeiras. Diferente do modelo supervisionado onde treinamos o algoritmo com um conjunto de treino e depois validamos as respostas com um conjunto de teste.

Para este trabalho utilizamos o modelo não supervisionado por se encaixar no cenário onde, dado um conjunto de dados observados , não temos um conjunto , de respostas associado.

### Redução de Dimensionalidade

Um dos principais problemas ao analisar uma base de dados é sua dimensão, pois na maioria das vezes apresentam um número elevado de atributos. Por exemplo em uma expressão genética, cada gene é representado por milhares de atributos, em um cenário como esse dificilmente será possível utilizar qualquer modelo de aprendizado de máquina com uma quantidade tão grande de atributos (FACELI, LORENA, *et al.*, 2011, p. 46).

Segundo Faceli, Lorena, et al., se cada atributo de uma base de dado for considerada uma coordenada em um espaço -dimensional, onde é o número de atributos, o volume desse espaço cresce exponencialmente com a adição de novos atributos. Ainda segundo Faceli, Lorena, et al., para que esses dados se tornem legíveis e compreensíveis é necessário reduzir a dimensão da base de dados e uma forma de minimizar o impacto do problema da dimensionalidade é combinar ou eliminar parte dos atributos irrelevantes.

“Em muitos algoritmos de AM (Aprendizado de Máquina), para que os dados com um número elevado de atributos possam ser utilizados, a quantidade de atributos precisa ser reduzida. A redução do número de atributos pode ainda melhorar o desempenho do modelo induzido, reduzir seu custo computacional e tornar os resultados obtidos mais compreensíveis.” (FACELI, LORENA, *et al.*, 2011, p. 46).

Nos próximos tópicos serão apresentados as técnicas utilizadas para solucionar o problema de redução de dimensão em nossa base de dados.

#### Técnica de Agregação

Para solucionar os problemas de muitas dimensões em uma base de dados surgiram diversas técnicas nas áreas de pesquisa como Reconhecimento de Padrões, Estatística e Teoria da Informação (FACELI, LORENA, *et al.*, 2011). Essas técnicas podem ser divididas em duas categorias:

* Agregação;
* Seleção de Atributos;

Segundo Faceli, Lorena, et al. (2011), a técnica de seleção de atributos é realizada em aplicações onde o número de atributos na base de dados sejam muito grandes, como a nossa base de dados é composta por somente 6 atributos esta técnica não é aplicável, por este motivo foi escolhido utilizar a técnica de Agregação para reduzir o número de dimensões em nossa base de dados. Por isso iremos detalha-lo melhor na próxima seção.

#### Analise de Componentes Principais (PCA)

As técnicas que utilizam agregação para reduzir a dimensão dos dados geralmente combinam os atributos originais por meio de funções lineares ou não lineares, e dentro dessas técnicas uma das mais conhecidas é a de Análise de Componentes Principais (PCA, do inglês *Principal Component Analysis*) (FACELI, LORENA, *et al.*, 2011).

Os atributos de uma base de dados representam as características sendo analisadas, por exemplo, se estamos analisando uma base de dados que possuem informações de compras de um cliente, informações como idade, gasto, salário e escolaridade são denominados de atributos da base de dados.

Quando estamos manipulando uma base de dados com uma quantidade de dimensão muito grande, a técnica de PCA nos permite resumir em dimensões menores mas que ainda assim representem a mesma variância do dado original (JAMES, WITTEN, *et al.*, 2013, p. 374). Para este trabalho estamos utilizando uma base de dados com 6 atributos somente, mesmo sendo um número pequeno seria extremamente complexo analisar um gráfico com 6 dimensões ou 6-D. Para solucionar este problema será utilizado a técnica de PCA para reduzirmos a complexidade dos nossos dados a um número que seja fácil de ser interpretado e ainda assim mantendo as informações importantes de nossa base de dados.

A primeira dimensão do PCA (PCA 1) é aquela onde a variação dos dados seja maior, por exemplo considere a Figura 4, a primeira dimensão é representada pela linha verde, pois neste gráfico os dados variam mais da esquerda inferior para a direita superior. A segunda dimensão do PCA (PCA 2) é representado pela linha tracejada azul, pois é a direção onde os dados tem a segunda maior variação (JAMES, WITTEN, *et al.*, 2013, p. 231).

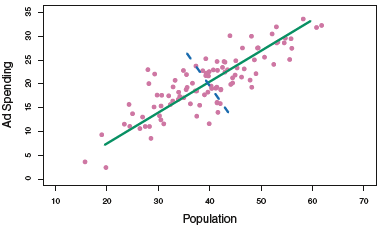


Figura : Funcionamento da Análise de Componentes Principais (PCA). Fonte: (JAMES, WITTEN, *et al.*, 2013, p. 230)

Segundo James, Witten, et al. (2013), se projetarmos todos os pontos de observação da Figura 4 na linha verde, iriamos obter a maior variância possível como pode ser observado na Figura 5. Como pode ser observado (Figura 5), a variância de cada dimensão de uma base de dado é calculada com base na direção onde o gráfico tem a maior variação. No caso da Figura 5 a primeira dimensão, é a variação dos dados horizontalmente, a segunda dimensão dos dados é a variação dos dados vertical.

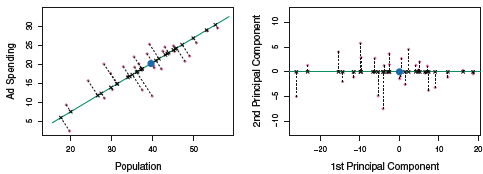


Figura : Projeção dos pontos para descobrir a maior variância. Fonte: (JAMES, WITTEN, *et al.*, 2013, p. 232)

Por se tratar de uma técnica que combina os valores dos atributos, isso leva a perda dos valores originais, o que pode ser um problema em alguns casos, por esse motivo é sempre bom manter os dados originais. Dessa forma ao obter o resultado final será possível interpreta-lo, associando o resultado produzido com os dados originais (FACELI, LORENA, *et al.*, 2011, p. 47).

Segundo James, Witten, et al., o objetivo do PCA é encontrar a menor representação -dimensional possível e ainda assim mantendo a maior quantidade de informação possível.

### Algoritmo de Agrupamento (Clustering)

O método de agrupamento (do inglês *Clustering*) tem como seu objetivo principal procurar por subgrupos ou “agrupamentos” dentro de uma base de dados. Ao encontrar esses subgrupos esperamos que cada dado dentro de um grupo sejam similares entre si, e que cada grupo seja diferente um do outro. (JAMES, WITTEN, *et al.*, 2013). Definir o que torna cada indivíduo do grupo similar ou cada grupo diferente um do outro depende muito do contexto em que esses dados estão sendo analisados.

Um exemplo de aplicação que utiliza o algoritmo de agrupamento é a segmentação de supermercados. Por exemplo, dado um conjunto de observações, cada um com atributos, sendo que o conjunto pode ser definido como clientes do mercado e sendo as características de cada cliente (ganho mensal, profissão, distância do mercado). O objetivo é criar subgrupos de pessoas que sejam mais receptíveis a propagandas de certos produtos, ou grupos de pessoas que tem maior probabilidade de realizar certos tipos de compras. (JAMES, WITTEN, *et al.*, 2013).

Devido a sua versatilidade, existem muitos métodos de *clustering,* na próxima seção entraremos em mais detalhes no método *clustering* chamado *K-means.*

### Algoritmo de Agrupamento: K-Means

Este algoritmo utiliza o conceito da distância Euclidiana para calcular as distâncias entre cada ponto de observação e os centros de cada grupo (). A distância euclidiana entre dois pontos e é definida pela fórmula:

Os grupos são definidos pelo algoritmo com base na menor distância entre os pontos de observações e os centros definidos. Em outras palavras, um ponto vai pertencer ao grupo se a distância entre eles for a menor comparada com a distância dos outros grupos ().

Na Figura 6 é possível observar o resultado do algoritmo K-means com 150 pontos de observações, utilizando diferentes valores para , sendo equivalente a um subgrupo.

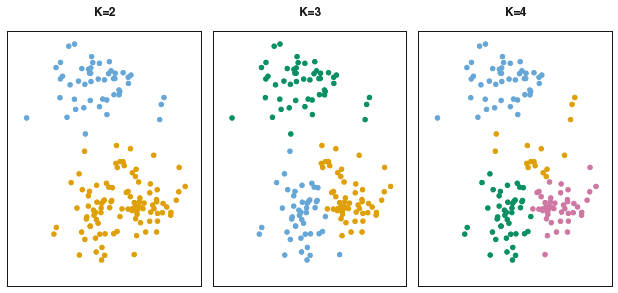


Figura : Exemplo do algoritmo K-means. Fonte: (JAMES, WITTEN, *et al.*, 2013, p. 387)

Segundo James, Witten et. al., uma das decisões mais importante do algoritmo K-Means é definir o número de centroides, pois não existe nenhuma regra de como esse valor é definido. Esta decisão pode variar de acordo com a necessidade do problema. Porém existem alguns métodos conhecidos que auxiliam na descoberta desse valor. Neste projeto serão utilizadas duas técnicas conhecidas como o método do cotovelo (Elbow Method) e o método do coeficiente de silhueta. Ambos serão tratados com mais detalhe na seção de desenvolvimento do projeto.

### Funcionamento do algoritmo K-means

Nesta seção será explicado qual o processo por trás do algoritmo K-means. Cada etapa será dividida em um item:

1. Definir o número de centroides () ou subgrupos que queremos dividir nossa base de dados;
2. Selecionar um ponto, dentro da base de dados onde será posicionado cada centroide ();
3. Atribuir cada ponto de observação ao centroide () mais próximo;
4. Calcular e centralizar a centroide de cada grupo;
5. Atribuir novamente cada ponto de observação ao centroide () mais próximo;
6. Repetir o passo 4 e 5 até que todos os pontos de observações estejam agrupados em seus respectivos centroides ();

Na Figura 7 fica claro como esse processo funciona passo a passo.

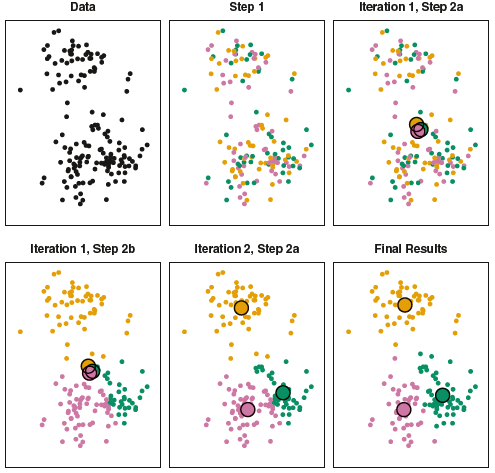


Figura : Processo do algoritmo k-means com k=3. Fonte: (JAMES, WITTEN, *et al.*, 2013, p. 389)

# Desenvolvimento

## Python

Para este trabalho estamos utilizando a linguagem de programação Python (Python Software Foundation, 2001-2017), na versão 3.6, a decisão de utilizar essa linguagem é devido ao fato de que ela já oferece muitas bibliotecas que auxiliam na criação de modelos de aprendizado de máquina. Nas próximas seções listaremos as bibliotecas utilizadas para este projeto.

## Base de dados

Os dados relacionados aos clientes foram extraídos do site UCI (Wholesale customers Data Set, 2007), onde estão disponíveis diversas fonte de dados para fins acadêmicos e de estudos sobre o tópico de aprendizagem de máquina.

A base de dados é composta por 6 categorias de produtos, onde cada linha representa uma transação realizada por um cliente. Cada cliente carrega a informação sobre quanto gastou em cada uma das 6 categorias de produtos. Com a base de dado definida, será possível avaliar quais informações serão importantes para o nosso modelo de segmentação, as informações que forem consideradas irrelevantes para o modelo serão descartados.

Os dados que forem considerados relevantes para o projeto será submetido ao modelo de segmentação para treinar o algoritmo, após o treinamento será possível segmentar os perfis dos clientes do atacado com base nas compras realizadas. As informações disponíveis são:

* Fresh: Gastos anuais com produtos frescos;
* Milk: Gastos anuais com produtos derivados de leite;
* Grocery: Gastos anuais com produtos comestíveis;
* Frozen: Gastos anuais com produtos congelados;
* Detergents\_paper: Gastos anuais com produtos de limpeza;
* Delicatessen: Gastos anuais com especiarias;
* Chanel: Tipo de estabelecimento do cliente (Hotel/Restaurante/Café);
* Region: Região do cliente;

Para este trabalho não usamos os dados Chanel e nem Region.

## Observação dos Dados

De acordo com Berry e Linoff (2004), data mining é apenas uma ferramenta, não é suficiente saber como utilizar e sim entender como será utilizado. Por este motivo é extremamente importante entender a base de dados sendo analisada para montar um modelo de aprendizado eficiente. A base de dados é composta por 6 atributos: 'Fresh', 'Milk', 'Grocery', 'Frozen', 'Detergents\_Paper' e 'Delicassen'. Cada atributo representa uma categoria de produtos vendidos pela distribuidora. Na tabela da Figura 8 estão algumas estatísticas extraídas da base de dados:

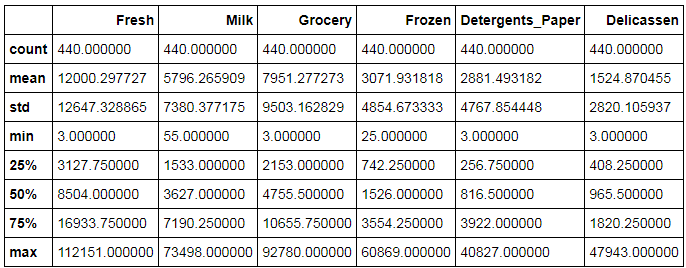


Figura : Descrição Estatísticas dos conjuntos de dados. Fonte Próprio Autor (2017)

### Análise dos Dados

Para ter melhor entendimento da base de dados, selecionamos uma amostra aleatoriamente para ser analisada separadamente. Os dados apresentados são referentes aos gastos anuais dos clientes, por este motivos os números representam unidades monetárias.

Utilizando a tabela da Figura 8 é possível realizar a análise do cliente da Figura 9, e definir um perfil com base nos gastos feitos em cada uma das categorias.

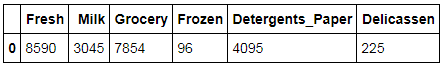


Figura : Amostra selecionada aleatoriamente. Fonte Próprio Autor (2017)

A principal compra do primeiro cliente é Fresh, com 8590. A média de compras nesta categoria é de 12000, ou seja, apesar de fresh ser sua principal compra ele está bem abaixo da média, está um pouco acima do segundo quartil que é 8504. Sua segunda maior compra é de Grocery, com 7854. A média de compras nesta categoria é de 7951, ele está próximo da média de compradores desta categoria, ele pertence ao segundo quartil com valor de 4755. Em terceiro lugar é Detergents\_Paper, com 4095. A média de compras nesta categoria é de 2881, ele se encontra acima da média, e está bem acima do terceiro quartil que é 3922. Isso significa que apesar desta categoria não ser sua maior despesa ele é um dos clientes que tem mais gastos com esta categoria. Em quarto lugar está a categoria Milk, com 3045. Este valor está abaixo da média desta categoria que é 5796 e se encaixa no primeiro quartil de compradores. O atributo Delicatessen está em penúltimo lugar em suas compras, com o valor de 225, muito abaixo da média que é 1524 e dentro do primeiro quartil de compradores. Em último lugar está Frozen com um gasto de 96, muito abaixo da média e um pouco acima do mínimo gasto com nesta categoria, isso significa que este cliente quase não compra produtos desta categoria.

Analisando um pouco mais afundo, podemos observar a distribuição dos atributos em um gráfico de uma matriz de dispersão, com ela é capaz de descobrir rapidamente se existe alguma correlação entre as categorias de produtos (Figura 10). Para interpretar a matriz de dispersão é necessário primeiro localizar o atributo que deseja analisar e verificar qual sua relação com os demais atributos, se o gráfico de dispersão entre dois atributos formarem uma linearidade, isso significa que esses atributos são correlacionados, ou seja, um cliente que compra um item tem chance de comprar o outro também.

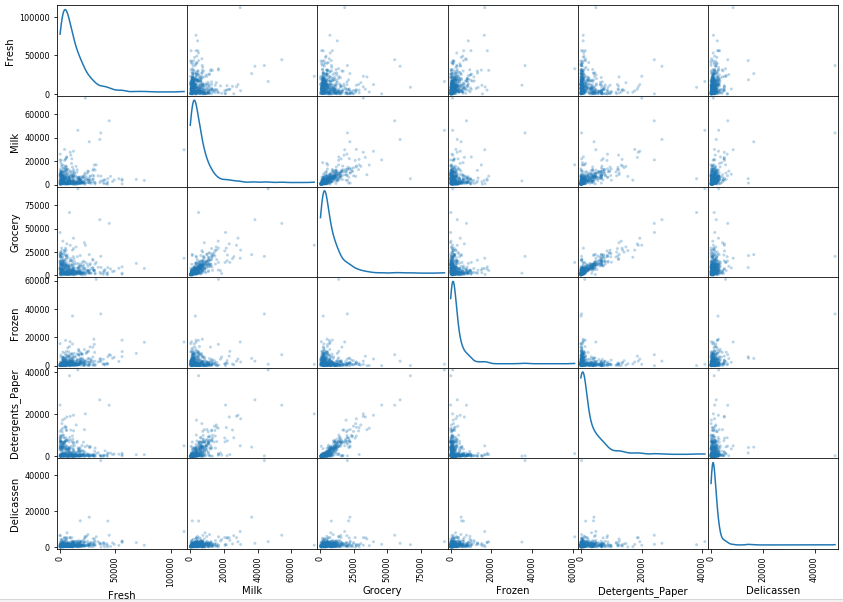


Figura : Matriz de dispersão, mostrando a correlação entre os atributos. Fonte Próprio Autor (2017)

Na matriz de dispersão (Figura 10) podemos notar que o gráfico das categorias "Detergents\_Paper" e "Grocery" formam uma linearidade, isso significa que há uma relação entre eles. Ou seja, um cliente que compre produtos da categoria "Detergent\_Paper" tem uma probabilidade maior de comprar produtos da categoria "Grocery" também. Existe uma relação parecida, porém pequena, entre as categorias "Grocery" e "Milk", mas a linearidade do gráfico não é totalmente visível.

Visualizando a Figura 10 não é possível ver essa correlação entre os atributos de forma nítida, porém se extrairmos outras amostras o gráfico de dispersão começa a fazer mais sentido. Na Figura 11, foram selecionados algumas amostras de clientes para mostrar a relação entre os atributos “Detergents\_Paper” e “Grocery”.

Com base nessas amostras selecionadas na Figura 11 é possível verificar que o que mostra no gráfico de dispersão é verdade. As três primeiras amostras mostram uma correlação entre os atributos “Detergents\_Paper” e “Grocery”, os gastos nestas duas categorias são aproximados. A mesma situação acontece com as últimas três amostras, onde é possível ver a relação entre os atributos “Milk” e “Grocery”. Isso significa que as essas categorias dividem uma relação mútua, ou seja, a probabilidade de um cliente comprar estes produtos em conjuntos é muito relativamente grande.

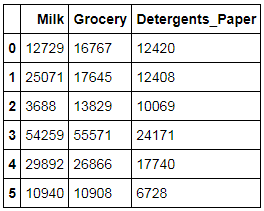


Figura : Amostras selecionadas para análise. Fonte Próprio Autor (2017)

### Resultado da análise dos dados

Com base nas análises realizadas no tópico anterior é possível deduzir algumas informações importantes sobre este cliente.

O primeiro ponto a ser notado é que se trata de um cliente de pequeno porte, isso é possível notar pela sua compra relativamente balanceada, indicando que ou sua clientela é muito pequena e não há necessidade ter um estoque muito grande, ou seu estabelecimento é pequeno e por conta disso seu estoque também.

O segundo ponto a ser observado é que sua principal fonte de renda é possivelmente a venda de produtos frescos, que podem variar de frutas, legumes ou pães. Na Figura 9 é possível notar que seu gasto com essa categoria é o mais alto em relação as demais categorias. Isso indica que os produtos frescos tem uma rotatividade maior dentro da loja, sendo necessário estocar e repor mais vezes.

O terceiro ponto a ser observado é que definitivamente a prioridade em seu estabelecimento não é vender produtos congelados, pois seu gasto nesta categoria é extremamente baixo, ou seja ele vende muito pouco produtos congelados e não há a necessidade de repor o estoque em grandes quantidades.

Quarto e último ponto é o fato de suas compras estarem, no geral, abaixo da média em todas as categorias, indicando que seu estabelecimento ou clientela são pequenos por isso não há a necessidade de repor seus produtos em grandes quantidades.

Com base nesses pontos levantados e nos dados analisados nos tópicos anteriores tudo indica que este cliente tem uma grande chance de ser um pequeno mercado de bairro ou uma loja de conveniência.

Este foi um exemplo de como seria feito a análise de um cliente para definir seu perfil e segmentar em grupos distintos, neste caso analisamos somente um cliente e rotulamos ele em 'loja de conveniência', com base em seus gastos nas 6 categorias. Esta análise foi feita sem a utilização de nenhum algoritmo ou código, essas informações foram extraídas apenas olhando os fatos, presentes nas amostras e gráficos. Porém nesta base de dados temos no total 440 amostras (Wholesale customers Data Set, 2007) e fazer esta análise manualmente um por um é tedioso e pode gerar margem ao erro. A proposta deste trabalho é utilizar aprendizado de máquina para realizar está análise através do algoritmo de agrupamento k-means.

## Pré-Processamento dos dados

### Escalonamento dos dados

Como foi explicado no tópico **2.1.2.1 Escalonamento dos dados**, utilizamos o logaritmo natural para escalonar os dados. Na Figura 12 é possível observar o resultado obtido.

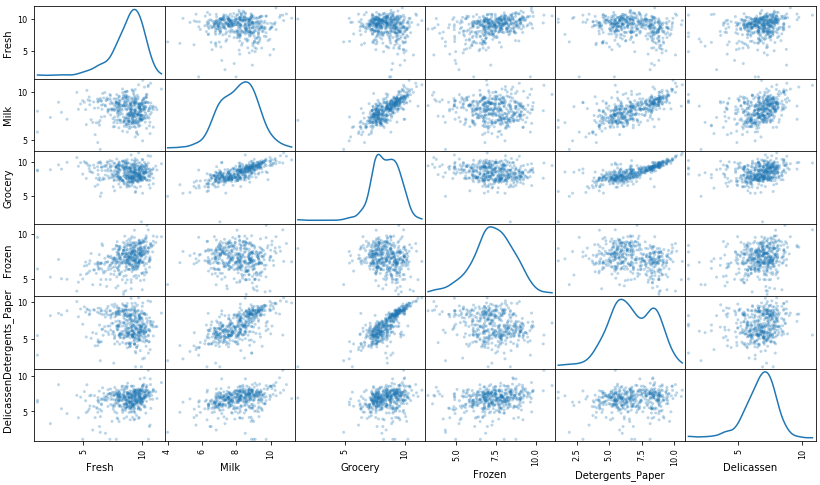


Figura : Matriz de dispersão após aplicar o logaritmo natural. Fonte Próprio Autor (2017)

Comparando com a Figura 10, onde os dados ainda não haviam sido escalonados é possível notar a diferença na dispersão dos pontos de observação. Os dados escalonados (Figura 12) mostram um gráfico mais claro, onde é possível ver com mais clareza a correlação entre os atributos “Detergent\_Paper” e “Grocery”. A linearidade entre estes dois atributos é mais presente ao escalonar os dados.

Na Figura 13 é possível notar a diferença dos dados antes e depois de aplicar o logaritmo natural. Agora a base de dados está escalonada e pronta para seguir para a próxima etapa.

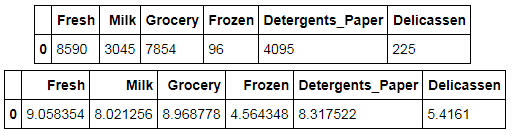


Figura 13: Redução da grandeza, após aplicar o logaritmo natural. A primeira tabela são os dados não escalonados, a segunda tabela são os dados escalonados. Fonte Próprio Autor (2017)

### Detecção de Desvios

O próximo passo para preparar a base de dados para o algoritmo K-means é detectar os pontos discrepantes na base de dados e remove-los. Para isso utilizaremos o método de Tukey, muito utilizado para identificar pontos de desvios.

Para trabalhar com o método de Tukey é necessário dividir a base de dados em 3 quartis , e . Cada quartil representa uma porcentagem do total da base de dados, e (SULLIVAN e LAMORTE, 2016). Segundo Sullivan e LaMorte (2016), um ponto de observação é considerado um desvio se:

Ou seja, de acordo com o método de Tukey, qualquer ponto que esteja abaixo de ou acima de pode ser considerado como desvio.

Após aplicar o método de Tukey foi possível encontrar diversos pontos de desvios em nossa base de dados, como pode ser observado na Figura 14.

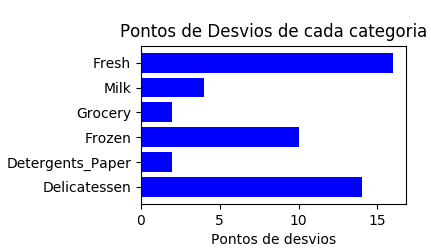


Figura 14: Pontos de desvios detectados em cada categoria. Fonte Próprio Autor (2017)

Segundo James, Witten, et al. (2013), o algoritmo K-means força cada ponto de observação a fazer parte de um grupo, durante a etapa de agrupamento. Por este motivo pontos de desvios podem causar impacto no resultado final se não forem tratados. Como a quantidade de pontos de desvios são relativamente pequenos comparado com o tamanho total da base de dados, optamos por remover esses pontos de desvios.

## Transformação dos Atributos

Nesta etapa do desenvolvimento iremos aplicar a técnica de PCA para reduzir a dimensão de nossos atributos, como estamos trabalhando com 6 atributos, seria muito complexo visualizar e analisar um gráfico 6-D.

O primeiro passo é aplicar o algoritmo de PCA e verificar quais dimensões tem maior variância. A variância de cada dimensão significa o quanto essa dimensão carrega de informação sobre a base de dados. A forma como a variância é extraída da base de dados está explicado com mais detalhes na seção Analise de Componentes Principais (PCA)

A Figura 15 mostra a variância de cada dimensão, agora que temos o resultado de quanto cada dimensão carrega de informação, vamos avaliar separadamente algumas dessas dimensões.

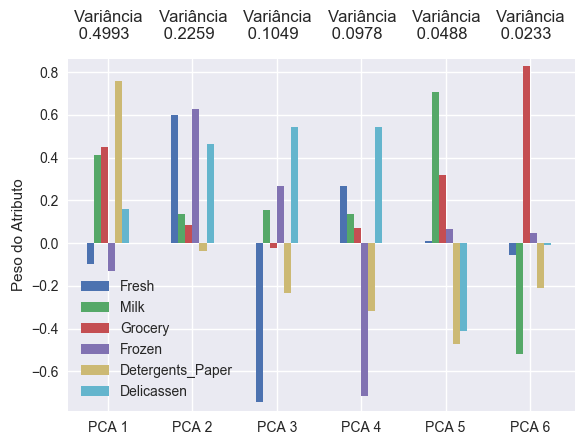


Figura 15: Variância de cada dimensão do PCA. Fonte Próprio Autor (2017)

### Análise das dimensões do PCA

Tomando como base a Figura 15, podemos analisar cada uma das dimensões geradas pelo PCA. Nesta seção iremos analisar algumas dessas dimensões para no final decidirmos quais dimensões utilizar para aplicar ao algoritmo de agrupamento K-means.

A somatória das variâncias de cada dimensão deve resultar em 1, isso significa que o conjunto das 6 dimensões presentes na Figura 15 carregam 100% das informações presente em nossa base de dados. A intenção nesta seção é conseguir o menor número de dimensões possíveis, mas que ainda assim carreguem informações suficientes para a nossa análise.

Na primeira dimensão (PCA 1), obtemos uma variância de 0.4993 ou seja, só a primeira dimensão carrega 50% das informações da nossa base de dados. O gráfico mostra uma forte relação entre os atributos ‘Detergents\_Paper, ‘Milk’ e ‘Grocery’, pois de acordo com o gráfico o peso desses atributos são próximos. Os atributos ‘Frozen’ e ‘Fresh’ mostram uma correlação negativa. Esta primeira dimensão consegue descrever bem os clientes que focam suas compras em ‘Detergents\_Paper’, ‘Milk’ e ‘Grocery’. Podemos deduzir que estes clientes sejam mercados de bairros, que foca mais suas vendas em materiais utilizados no dia a dia.

Na segunda dimensão (PCA 2), obtemos uma variância de 0.2259, ou seja, ela carrega 20% das informações da nossa base de dados. Podemos observar uma forte relação entre 3 atributos, ‘Fresh’, ‘Frozen’ e ‘Delicatessen’. O gráfico mostra também uma relação pequena, porém negativa com o atributo ‘Detergents\_Paper’. Poremos deduzir que estes clientes sejam restaurantes, já que suas compras são focadas em alimentos frescos, congelados e especiarias.

Na terceira dimensão (PCA 3), obtemos uma variância de 0.1049, ou seja, ele carrega aproximadamente 10% das informações da nossa base de dados. Nesta dimensão existe uma relação positiva entre os atributos ‘Delicatessen’, ‘Frozen’ e ‘Milk’ e uma relação extremamente negativa em ‘Fresh’ e um pouco negativa em ‘Detergents\_Paper’. A partir destas informações podemos deduzir que estes clientes sejam lojas de conveniências, onde claramente seu foco não é vender produtos frescos mas sim produtos congelados, especiarias e derivados de leite.

### Conclusão da análise das dimensões

Como pode ser observado na Figura 15 a tendência é que quanto maior a dimensão menor a variância, ou seja a relevância das informações carregadas vão diminuindo ao longo das dimensões geradas pelo PCA.

A quarta, quinta e sexta dimensão, por exemplo, carregam no total aproximadamente 17% da nossa base de dados, ou seja, são dimensões que podem ser descartadas, pois suas variâncias são muito baixas, ou seja elas explicam muito pouco sobre a nossa base de dados. Por este motivo para este trabalho não vamos utilizar as 3 últimas dimensões. Agora nos restam as três primeira dimensões, que no total explicam 0.8301 ou aproximadamente 83% da nossa base de dados. O que é um valor significativo e muito bom.

Porém se utilizarmos as 3 dimensões ainda restaria um gráfico 3-D para ser analisado, para simplificar mais resolvemos remover também a terceira dimensão, apesar dela representar aproximadamente 10% dos dados será muito mais prático para fins acadêmicos analisar um gráfico 2-D.

Segundo Géron (2017) é favorável que seja selecionado um número de dimensões que na totalidade somem pelo menos 95% da variância, a menos que seja necessário a visualização gráfica, neste caso é recomendado que reduza as dimensões para 2 ou 3.

Então finalmente chegamos à conclusão que a primeira e a segunda dimensão são as que mais explicam nossa base de dados, no total as duas dimensões apresentam 0.7252 de variância ou aproximadamente 72%. Por este motivo utilizaremos as duas primeiras dimensões para aplicar o algoritmo de agrupamento K-means.

## Algoritmo de agrupamento (K-Means)

Depois de passar por todos os passos de analisar a base de dados, escalonar os dados, remover os desvios e reduzir a dimensão da base de dados, finalmente podemos dar início ao modelo de aprendizado de máquina K-Means.

O primeiro passo para aplicar o algoritmo de K-means é identificar a quantidade grupos () que desejamos segmentar a nossa base de dados. No nosso caso cada grupo irá representar um perfil de cliente. O número de grupos selecionados pode ter um grande impacto no resultado final, geralmente o ideal é tentar várias combinações e analisar qual resultado gera uma melhor interpretação dos dados. Não há uma resposta certa de quantos grupos () um algoritmo de K-Means deve ter (JAMES, WITTEN, *et al.*, 2013, p. 400).

Porém existem alguns métodos que nos auxiliam a descobrir um número próximo do ideal para os clusters, algum deles serão apresentados a seguir. Durante as próximas seções estaremos referenciando cada grupo como *centroides* ou *clusters*

### Método do cotovelo (Elbow Method)

A ideia do Elbow Method é definir os clusters onde a variação dos pontos de observação de cada cluster seja a menor possível, essa variação é definida pela soma dos clusters internos ao quadrado, ou em inglês (*Within-cluster Sum of Square (WSS)*) (KASSAMBARA, 2017, p. 129). Ou seja, para cada cluster é calculado a distância de todos os pontos que pertencem a ele, e a partir disso analisar se ao adicionar mais um cluster a variância dessa somatória aumenta ou diminui. Se ao adicionar mais cluster essa variância diminuir, então temos o número ideal de clusters.

Na Figura 16 é possível observar a variação do WSS em relação aos clusters. A partir do gráfico é possível deduzir que o número ideal de clusters está entre 2 e 4. A partir de 5 clusters é possível notar que a variação das somas das distâncias começam a diminuir muito.

Um cluster com a variância muito alta, significa que ele carrega muita informação, e isso não é bom pois não carrega uma característica específica, as mesma coisa acontece com um cluster com uma variância muito pequena, que significa que ele carrega tão pouca informação que não possui uma característica específica também. O ideal é encontrar uma quantidade de clusters onde a variação seja mediana.

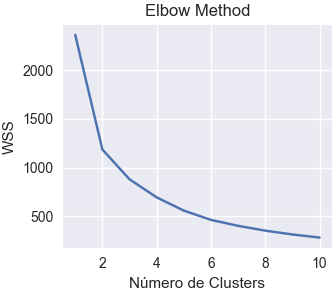


Figura 16: Método do cotovelo aplicado a base de dados dos clientes do atacado. Fonte Próprio Autor (2017)

A partir do método de Elbow conseguimos deduzir que o número ideal de clusters para o nosso modelo está em torno de 2 a 4 clusters. Mas precisamos de algo mais concreto para decidir quantos clusters utilizar no modelo K-Means, para isso iremos utilizar um outro método mais preciso para identificar os cluster na nossa base de dados.

### Método do coeficiente de silhueta

De acordo com Kassambara (2017), o método de silhueta calcula a qualidade de um cluster avaliando os pontos de observação presentes nele. Para cada cluster é atribuído uma média que condiz com sua qualidade, quando maior for sua média, melhor.

O cálculo do coeficiente de Silhueta é feito a partir da média das distâncias interna dos pontos de observação dentro de um cluster e a média da distância do cluster mais próximo (ROUSSEEUW, 1987, p. 55).

Onde, representa as distâncias internas de um cluster e a distância do cluster mais próximo. A Figura 17 exemplifica como o coeficiente é obtido.

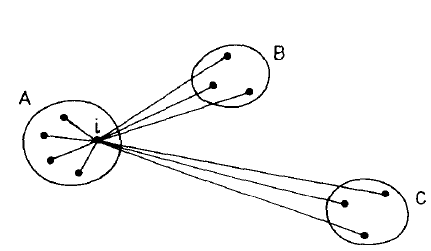


Figura : Ilustração dos elementos envolvidos no cálculo do coeficiente de Silhueta. Fonte: (ROUSSEEUW, 1987, p. 55)

Para o nosso trabalho simulamos o coeficiente de silhueta para 10 clusters, assim como foi feito com o método de Elbow

A Figura 18 mostra o resultado obtido a partir do cálculo do coeficiente de Silhueta. Como pode ser observado a melhor média que o algoritmo computou foram com 2 clusters. Se voltarmos para o gráfico do método de Elbow (Figura 16), fica claro que a partir de 2 clusters a variação das distancias tem uma mudança súbita, o que prova que para o nosso modelo o número ideal de clusters será 2.

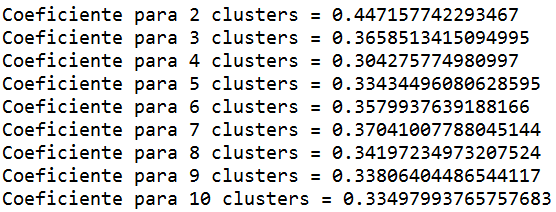


Figura 18: Resultado do coeficiente de silhueta na base de dados. Fonte Próprio Autor (2017)

### K-Means

Após definir o número de clusters que serão utilizado em nosso modelo, podemos finalmente aplicar o algoritmo de agrupamento K-Means para descobrir os diferentes agrupamentos para os clientes do atacado.

Na Figura 19 temos o resultado final do algoritmo de agrupamento K-Means, como pode ser observado os clientes foram divididos em 2 grupos (clusters) distintos. Agora que já temos definido os grupos dos clientes da nossa base de dados, precisamos analisar o que cada grupo representa.

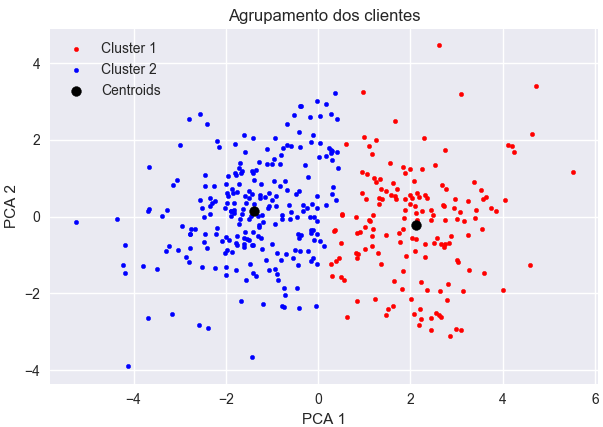


Figura 19: Resultado do algoritmo K-Means, na base de dados dos clientes do atacado. Fonte Próprio Autor (2017)

## Data Mining

Após aplicar o algoritmo de K-Means precisamos analisar o que cada cluster significa, para isso precisamos obter o valor original do centro de cada cluster e analisar separadamente.

Na Figura 20, podemos observar o valor original dos gastos dos clientes que representam o centro de cada cluster. Com base nesses valores podemos concluir o que cada cluster representa.

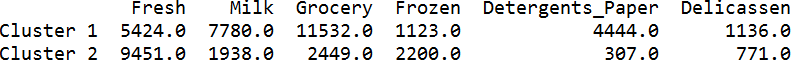


Figura 20: Valor original do centro de cada cluster. Fonte Próprio Autor (2017)

O principal ponto a se notar no cliente do cluster 1, é que exceto a categoria ‘Fresh’, em todas as outras categoria seus gastos são bem superiores ao cliente do cluster 2. Isso denota o perfil de um cliente que seus gastos são maiores, e se seus gastos são maiores isso significa que este cliente tem um número de clientela muito grande a ponto dele ter que repor seus produtos em grande quantidade, ou que seu estabelecimento é de grande porte, por isso em quase todas as categorias seus gastos são muito grande. A partir destas análises é possível deduzir que o cliente pode ser um supermercado ou um revendedor.

Já o cliente que pertence ao cluster 2 tem seu gasto focado em ‘Fresh’, todas as outras categorias apresentam um gasto relativamente baixo comparado ao cliente do cluster 1. Isto denota o perfil de um cliente que não é um revendedor, pois seus gastos são muito pequenos, e pelo fato de seu foco ser com produtos frescos este cliente possivelmente compra produtos para preparar refeições. Logo este cliente pode ser um restaurante ou lanchonetes, onde o foco não é revender produtos mas sim utiliza-los para a preparação de refeições que serão servidas no local.

# Conclusão

Este trabalho pode ser dividido em duas partes que foram extremamente importantes: a obtenção e o tratamento dos dados. Ambas as partes foram fundamentais para a conquista do objetivo deste projeto que é segmentar os clientes de um atacado com base nos seus gastos em diferentes categorias de produtos.

A fase da obtenção de dados apresentou dificuldades pois, não são todos os estabelecimento que estão dispostos a revelar informações confidenciais sobre seus modelo de negócios. As bases de dados encontradas durante a pesquisa ou eram simuladas, muito pequenas ou muito grandes, o que seria um problema para o projeto. As bases de dados simuladas podem gerar inconsistência nos dados, ou no pior dos casos mesmo aplicando um modelo de aprendizado não seria possível encontrar nenhum padrão, pois não condiz com um comportamento real de um cliente. Já as bases de dados que eram muito pequenas não possuíam dados suficientes para serem computados e extrair informações interessantes. Já as bases de dados grandes traziam outros problemas de complexidade, que vão além do escopo deste projeto.

Por fim conseguimos encontrar uma base de dados de um atacado disponível no site da UCI (Wholesale customers Data Set, 2007), um repositório de dados para estudos de aprendizado de máquina.

A fase do tratamento de dados foi extremamente importante, visto que grande parte do desenvolvimento do projeto se encontra nesta fase. A grande dificuldade durante o tratamento dos dados foi compreender e aplicar os métodos para processar os dados de maneira correta, para que pudesse ser consumido pelo algoritmo K-Means, pois se os dados não forem pré-processados corretamente não seria possível chegar a um resultado satisfatório.

Após a conclusão dessas duas partes conseguimos obter o resultado final e cumprir com o objetivo deste trabalho. Onde mostramos a segmentação dos clientes com base em seus gastos nas diversas categorias de um atacado.

# Bibliografia

BERRY, M. J. A.; LINOFF, G. S. **Data Mining Techniques For Marketing Sales And Customer Relationship Management 2Ed**. [S.l.]: Wiley, 2004.

CIL, I.; AY, D.; TURKAN, Y. S. **DATA DRIVEN DECISION SUPPORT TO SUPERMARKET LAYOUT**. [S.l.]: [s.n.], 2009.

FACELI, K. et al. **Inteligência artificial:** Uma abordagem de aprendizado de máquina. [S.l.]: LTC, 2011.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. **From Data Mining to Knowledge Discovery in Databases**. [S.l.]: [s.n.], 1996.

GÉRON, A. **Hands-On Machine Learning with Scikit-Learn and TensorFlow**. First Edition. ed. [S.l.]: O’Reilly Media, 2017.

HAMILTON, J. Use of logarithms in economics. **Econbrowser**, 2014. Disponivel em: <http://econbrowser.com/archives/2014/02/use-of-logarithms-in-economics>. Acesso em: 24 out. 2017.

JAMES, G. et al. **An Introduction to Statistical Learning**. New York: Springer, 2013.

KASSAMBARA, A. **Practical Guide to Cluster Analysis in R:** Unsupervised Machine Learning (Multivariate Analysis). 1ª. ed. [S.l.]: sthda, v. 1, 2017.

KHANAL, S. R. **Machine Learning for Supermarket Data Analysis**. [S.l.]: [s.n.], 2016.

OHTA, M.; HIGUCHI, Y. **Study on the Design of Supermarket Store Layouts:** The Principle of “Sales Magnet”. [S.l.]: [s.n.], 2013.

PRASS, F. Business Intelligence. **Business Intelligence**, 10 jul. 2017. Disponivel em: <http://fp2.com.br/blog/index.php/2012/um-visao-geral-sobre-fases-kdd/>.

PYTHON Software Foundation. **Python**, 2001-2017. Disponivel em: <https://www.python.org/>. Acesso em: 12 out. 2017.

ROUSSEEUW, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. **Journal of Computatuinal and Applied Mathematics**, Fribourg, 1987. 53-65.

SULLIVAN, L.; LAMORTE, W. W. InterQuartile Range (IQR). **Boston University School of Public Health**, 2016. Disponivel em: <http://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704\_summarizingdata/bs704\_summarizingdata7.html>. Acesso em: 28 out. 2017.

WHOLESALE customers Data Set. **UCI Machine Learning Repository**, 2007. Disponivel em: <https://archive.ics.uci.edu/ml/datasets/Wholesale+customers>. Acesso em: 12 out. 2010.

# APÊNCIDE A – Código fonte do desenvolvimento do agrupamento de perfis dos clientes

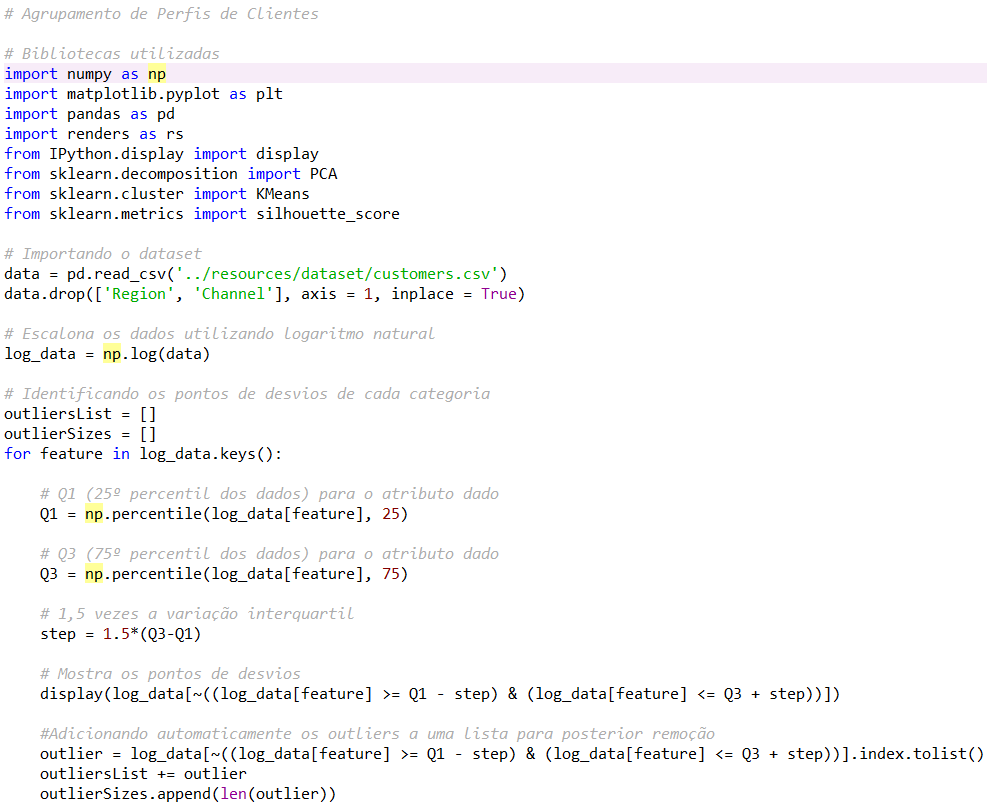


Figura : Trecho do código onde é identificado os pontos de desvios. Fonte: Próprio Autor (2017)



Figura : Trecho de código onde é aplicado o PCA. Fonte: Próprio Autor (2017)

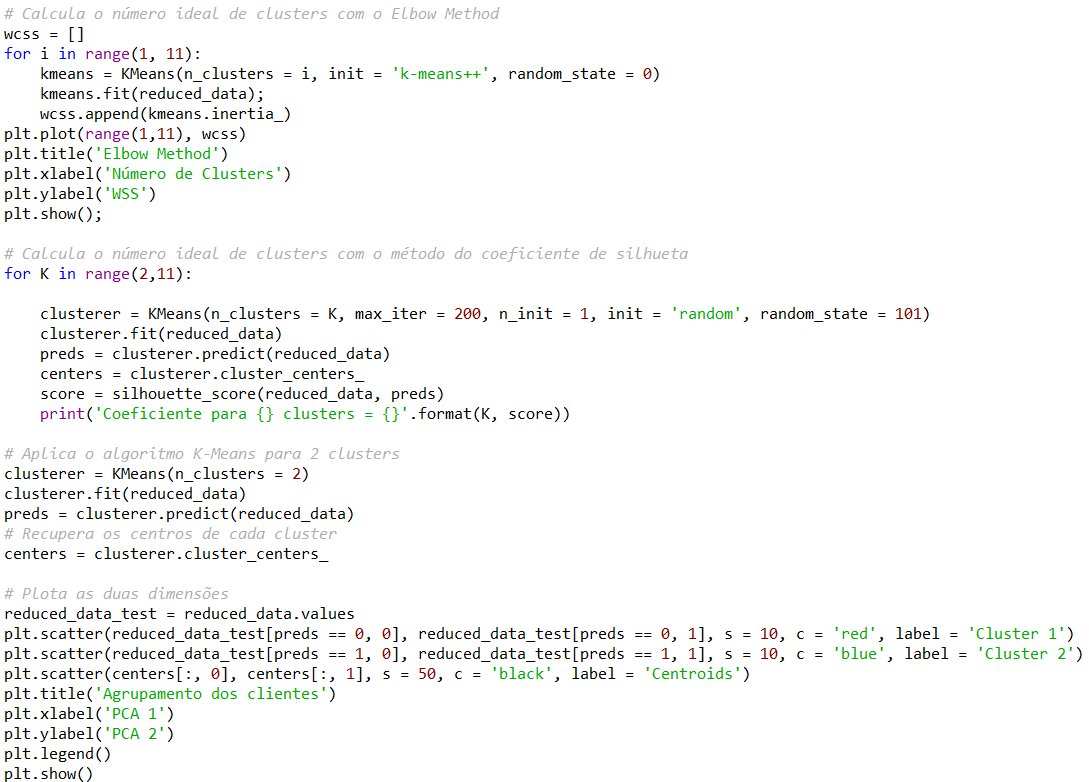


Figura : Trecho do código onde é aplicado o algoritmo K-Means. Fonte: Próprio Autor (2017)

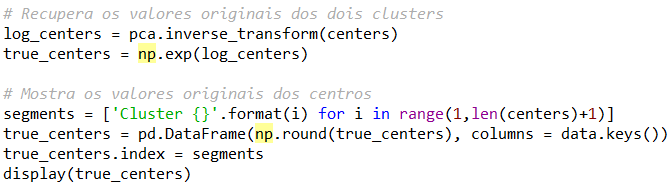


Figura : Trecho do código onde é recuperado os valores originais dos clusters. Fonte: Próprio Autor (2017)