The background is a dark blue-grey color with a complex, abstract pattern of thin, light-colored lines and small, multi-colored dots (yellow, blue, orange, and white). These elements form a network-like structure, reminiscent of a data visualization or a molecular model, with lines connecting various points across the frame.

# Big data e ciência de dados

## aula 06

# Análises não-supervisionadas

Luisa Mariele Strauss

Provost & Fawcett (2016)  
Material profa. Patrícia Kuyven

# Relembrando - Tipos de estudos



Observacional: entender, sem mudar



Experimento: mudar variáveis de parte da população



Simulação: modelagem para reproduzir parte da realidade



Análise de dados: vários níveis

# Tipos de análise de dados

- Exploratória vs explanatória
- Descritiva
- Diagnóstica
- Preditiva
- Prescritiva





# Tipos de problemas e caminhos de solução

---

- Quem são os clientes mais lucrativos?
- Quais as características dos clientes mais lucrativos?
- Será que um novo cliente em particular será lucrativo?
- Existe uma classificação para os meus produtos, além do nosso catálogo de produtos?
- Quais clientes tem gostos parecidos?



## Análises em ciência de dados e data mining

---

- Não supervisionada: quando não há um alvo
- Supervisionada: quando há um alvo a ser explicado





# Exemplos de problemas

Como podemos agrupar nossos clientes? **Não supervisionada**

- Não há uma variável alvo a ser atingida, usamos os atributos dos clientes
- Número de compras, valor das compras, tipos de itens, idade, profissão, escolaridade etc

Qual a chance deste cliente ser bom pagador? **Supervisionada**

- Alvo: pagou (sim/não) ou pagou em dia (sim/não)
- Demais atributos ajudam a explicar e/ou prever o alvo

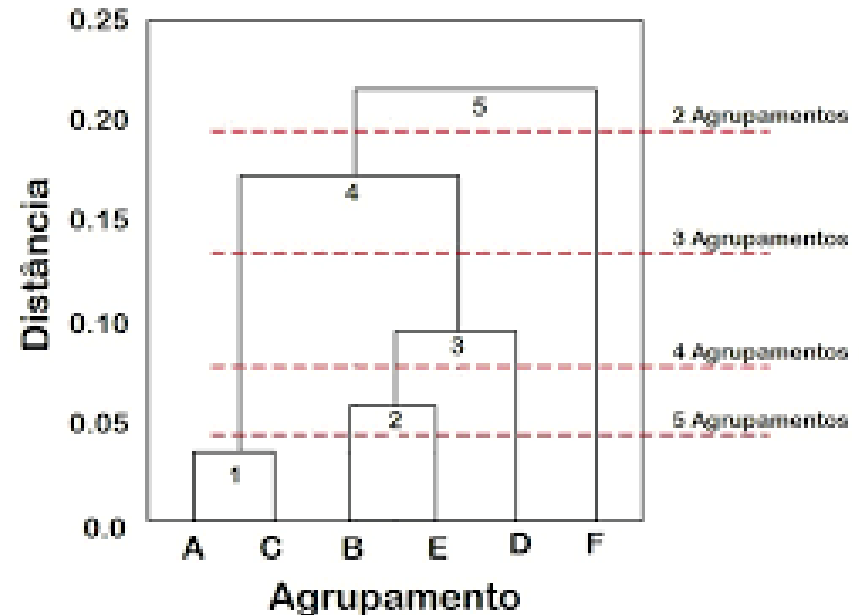
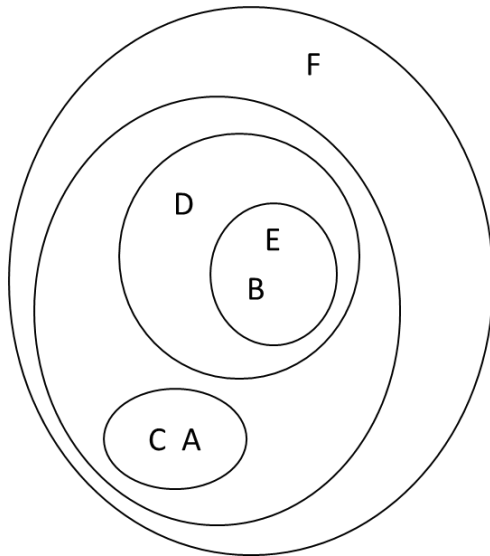
## Não supervisionada - Agrupamentos ou clusters

- Busca de grupos de indivíduos sem ter parâmetros prévios definidos
- Indivíduos: clientes, fornecedores, produtos, ...
- Pode ser hierárquica ou grupos exclusivos



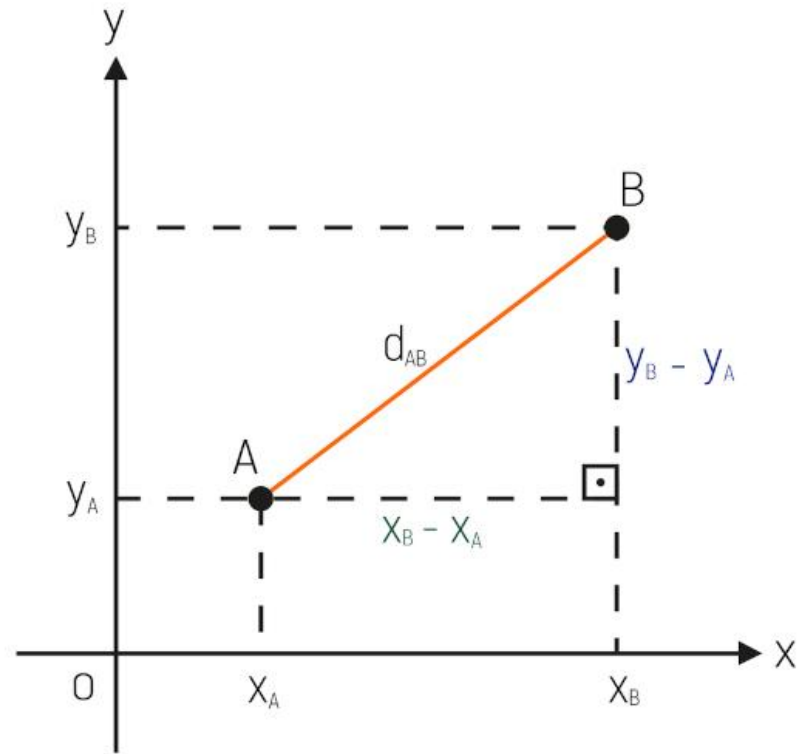
## Agrupamento hierárquico

- Mostra agrupamentos por semelhança, em que os indivíduos e grupos estão contidos em grupos maiores.
- Permite ver a “paisagem” dos dados.



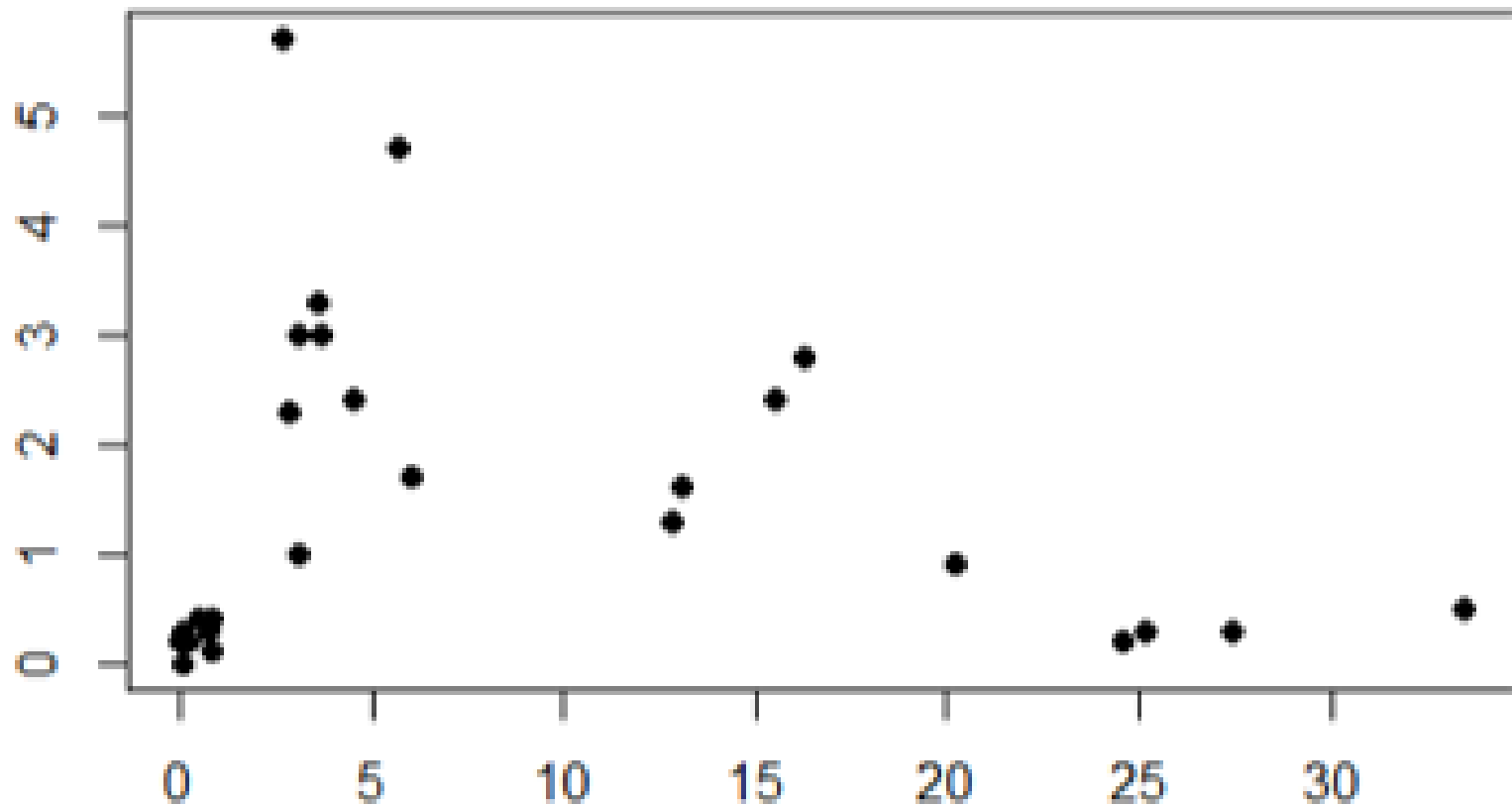


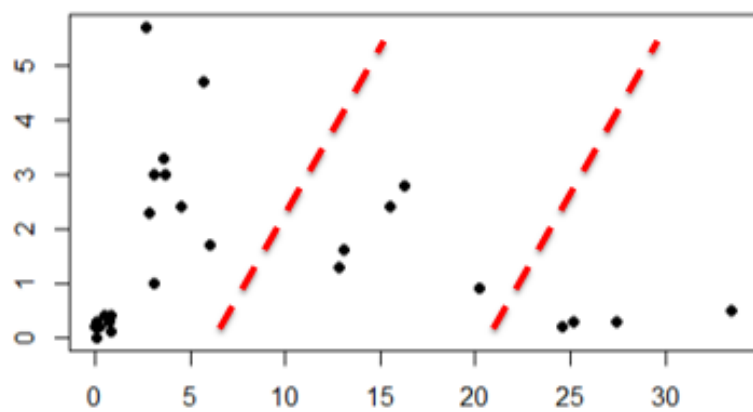
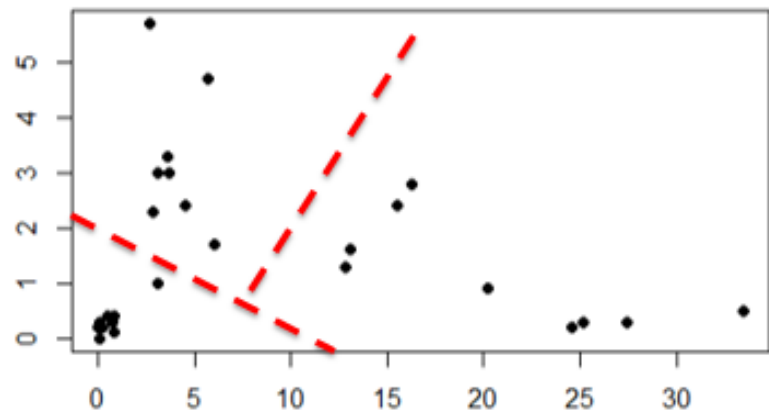
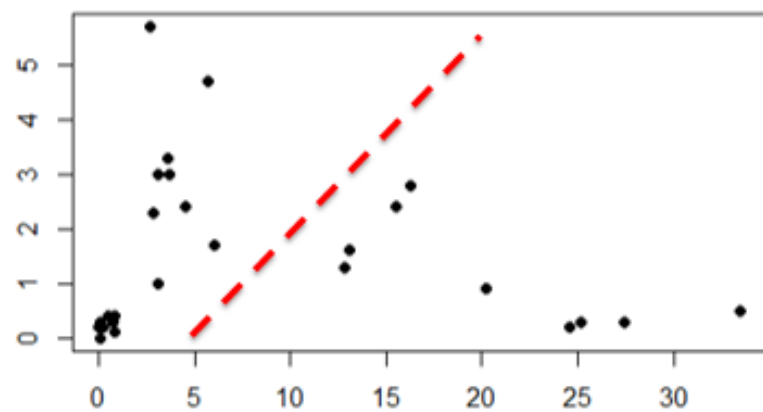
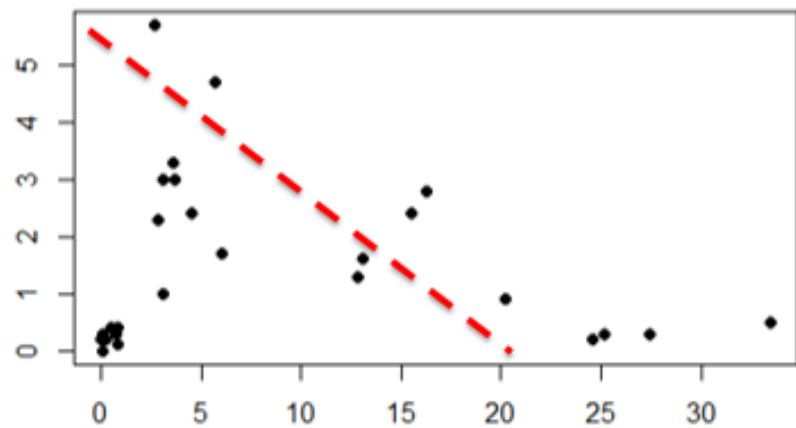
# Similaridade e distância



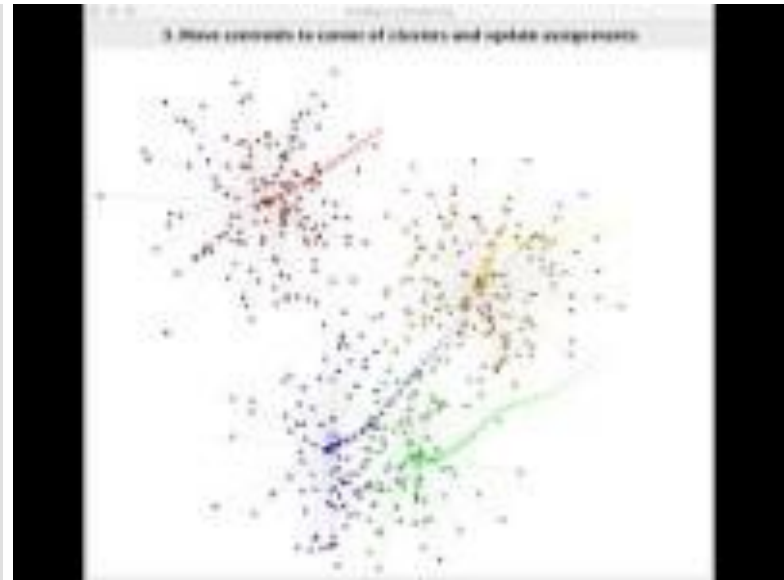
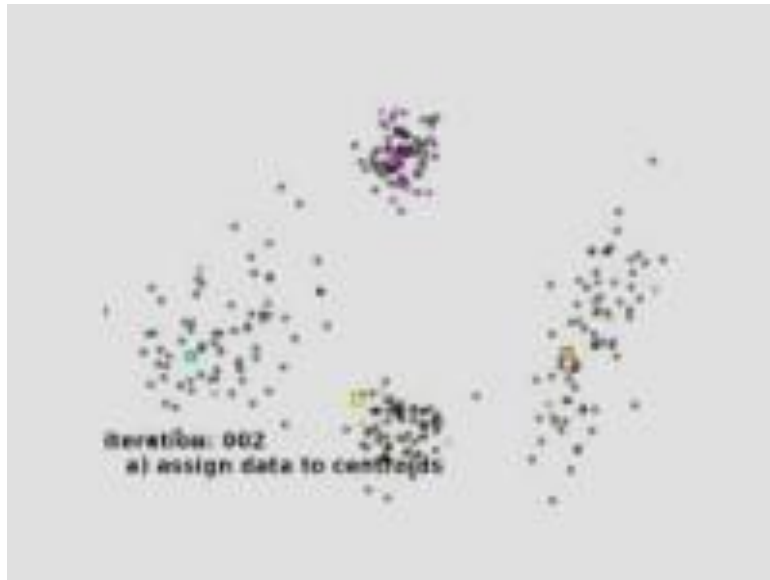
$$d_{AB} = \sqrt{[x_B - x_A]^2 + [y_B - y_A]^2}$$

Imagem obtida em <https://mundoeducacao.uol.com.br/matematica/distancia-entre-dois-pontos.htm>



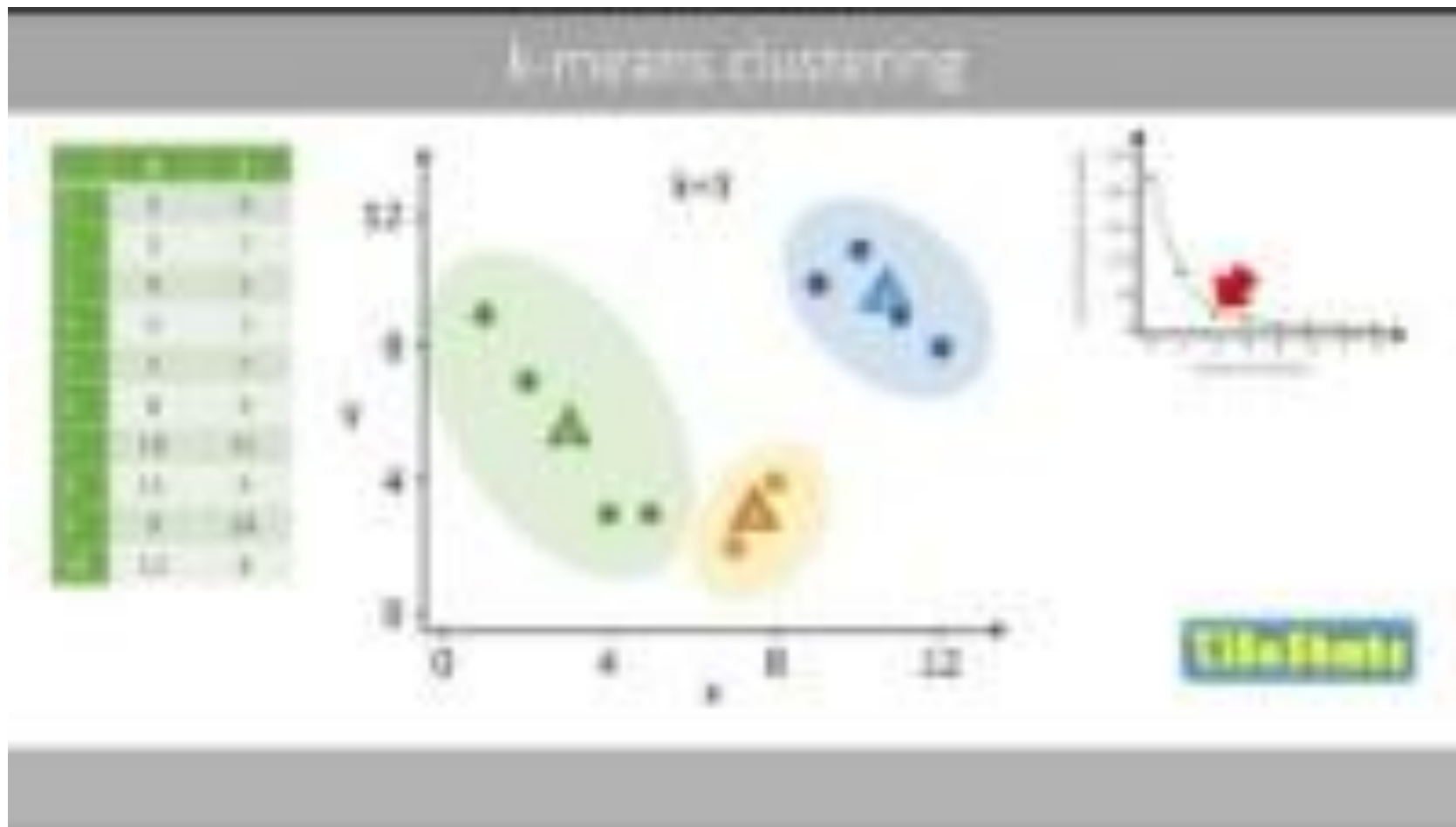


# Clustering (Método k-means)

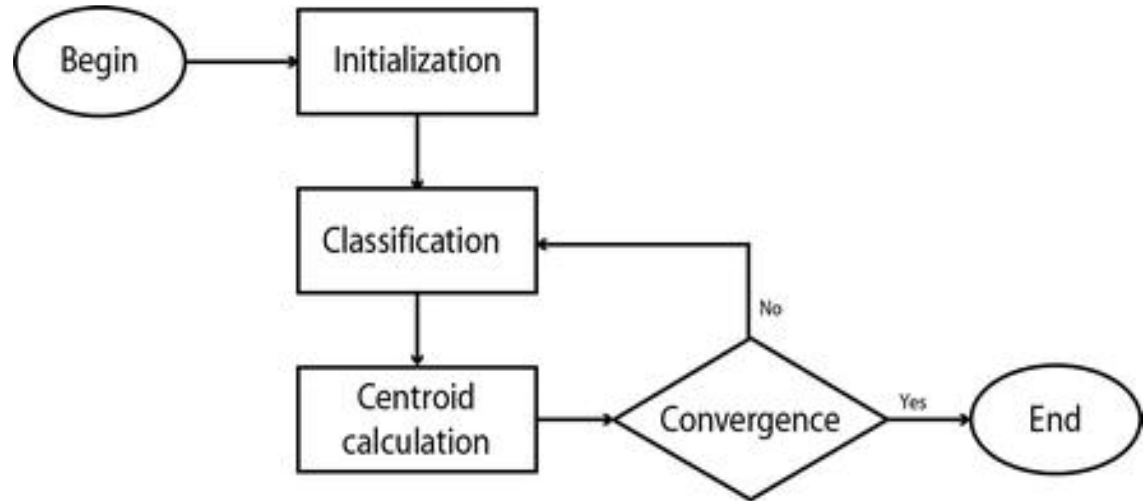





# Animação com explicação do método k-means



## Etapas do k-means





A base usada por  
vocês permite essa  
análise? É útil?



# Fazendo cluster no Power BI

- A partir de gráfico de dispersão
  - Método usado: Expectation-Maximization Clustering\*
  - Limitação: apenas duas variáveis numéricas
-





# Fazendo cluster com Phyton

- Método usado: K-means
  - Biblioteca para machine learning - K-means: scikit-learn ou sklearn
  - Outras bibliotecas: numpy, pandas e matplotlib
  - Importante: para k-means, as variáveis devem ser numéricas
-

# Etapas



Importar bibliotecas



Carregar o conjunto de dados



Limpar e transformar os dados



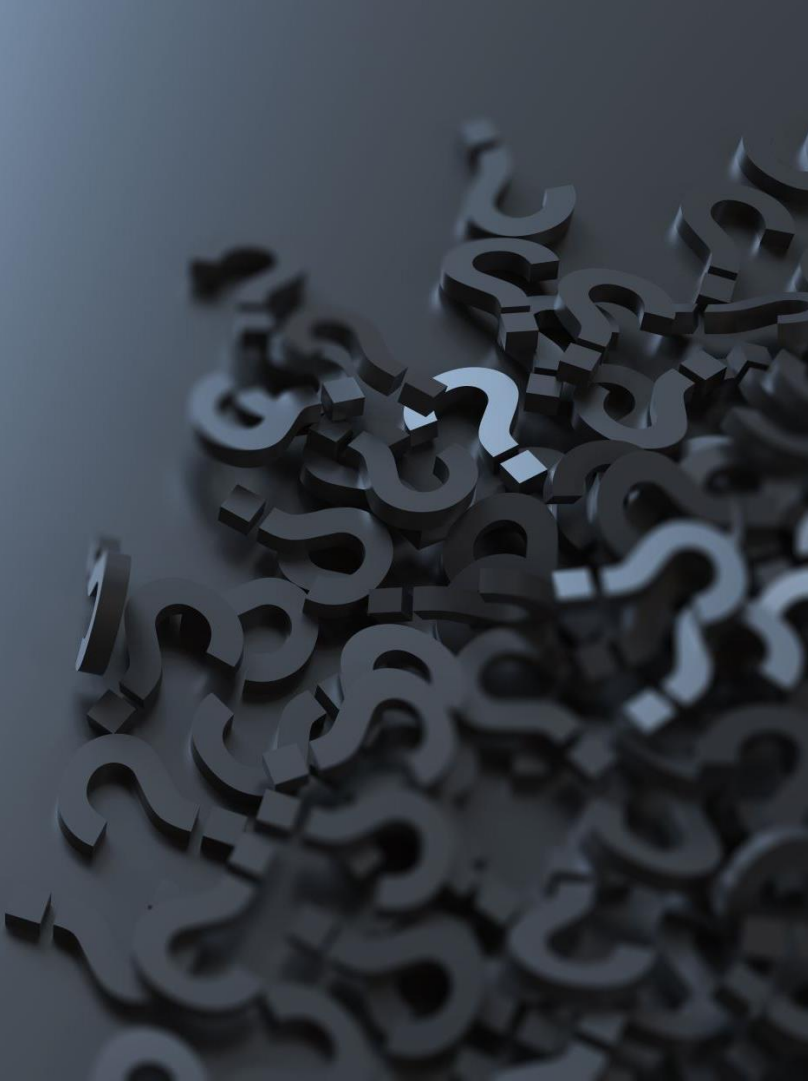
Encontrar o número de clusters



Implementar o k-means




Visualizar os clusters



---

## Considerações importantes

- A ideia é que os elementos de cada grupo tenham alta similaridade entre si, mas diferenças entre outros grupos
- Nem sempre é possível explicar facilmente os grupos gerados
- Então, como analisar e entender os grupos? Como saber se o modelo é adequado?



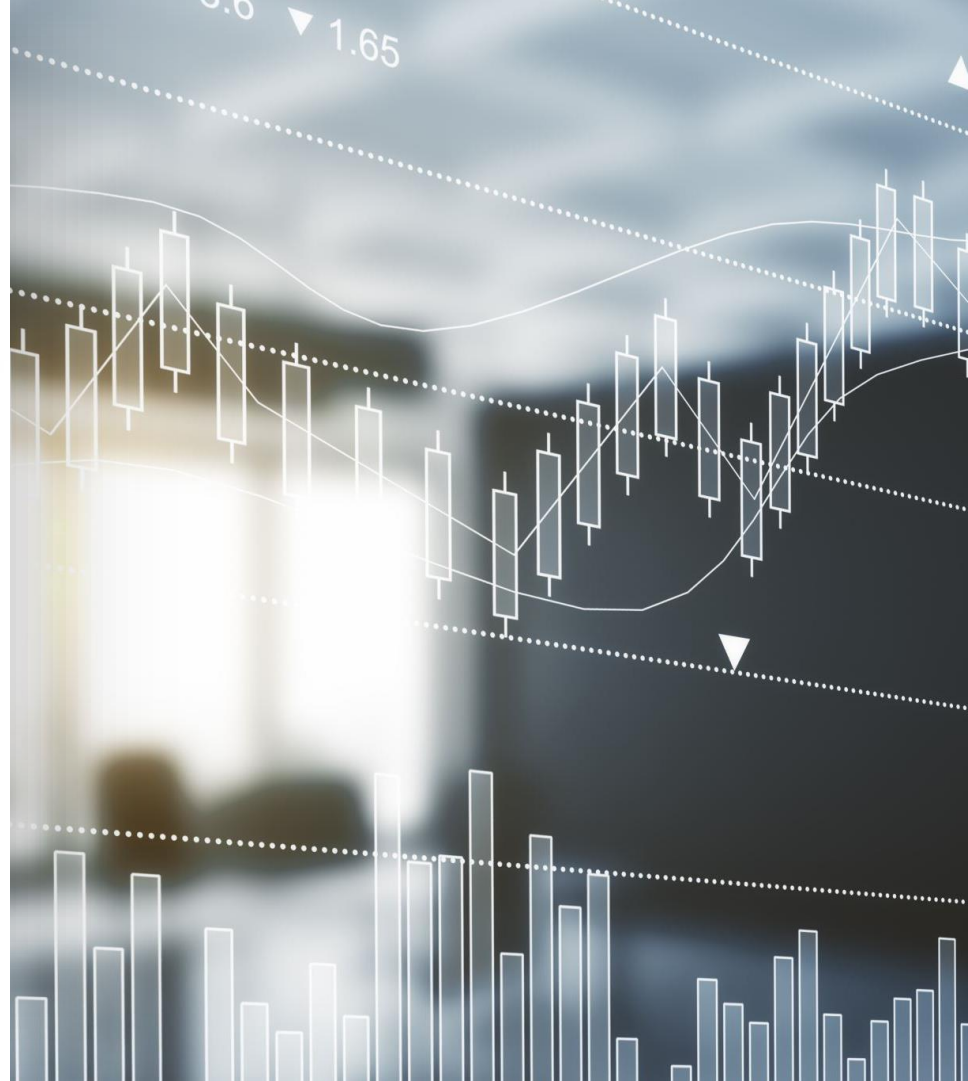
## Base para exercício - Wholesale customers

- Disponibilizada pela UCI em <https://archive.ics.uci.edu/dataset/292/wholesale+customers>
- Representa vendas de um atacadista, com os dados a seguir
  - Channel - canais de distribuição, sendo 1 - Horeca (Hotel/Restaurante/Café e 2 – Varejo
  - Region - região em que os clientes estão localizados, sendo 1 – Lisboa, 2 - Porto e 3 - Demais regiões
  - Fresh, Milk, Grocery, Frozen, Detergents\_paper, Delicatessen – cada variável representa o valor monetário gasto em cada categoria



# Normalização de dados

- Objetiva deixar os dados em uma mesma “escala”
- Normalização coloca os dados em um intervalo entre 0 e 1 ou -1 e 1
- Função `MaxAbsScaler` da Biblioteca `sklearn`



# Exercício

- Rodar o k-means para Wholesales\_customer sem normalizar e normalizando dados
- Rodar o k-means para Wholesales\_customer com diferentes números de clusters
- Interpretar cada agrupamento
- Escolher um agrupamento
- Rotular o agrupamento
- Levantar possibilidades de ação para cada agrupamento
- Inserir as variáveis Channel e Region e visualizar os diferentes agrupamentos

