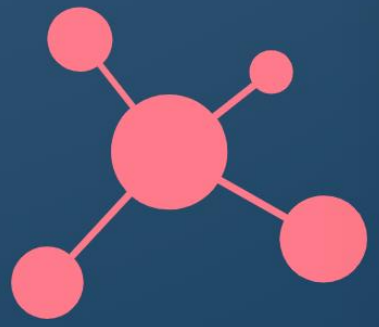
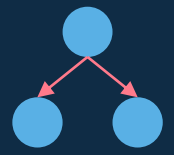
Tech&Dados
BY @ITALLEIRAS



Agrupamento: K-Means

Funcionamento, exemplos, código e mais.

Novembro de 2023

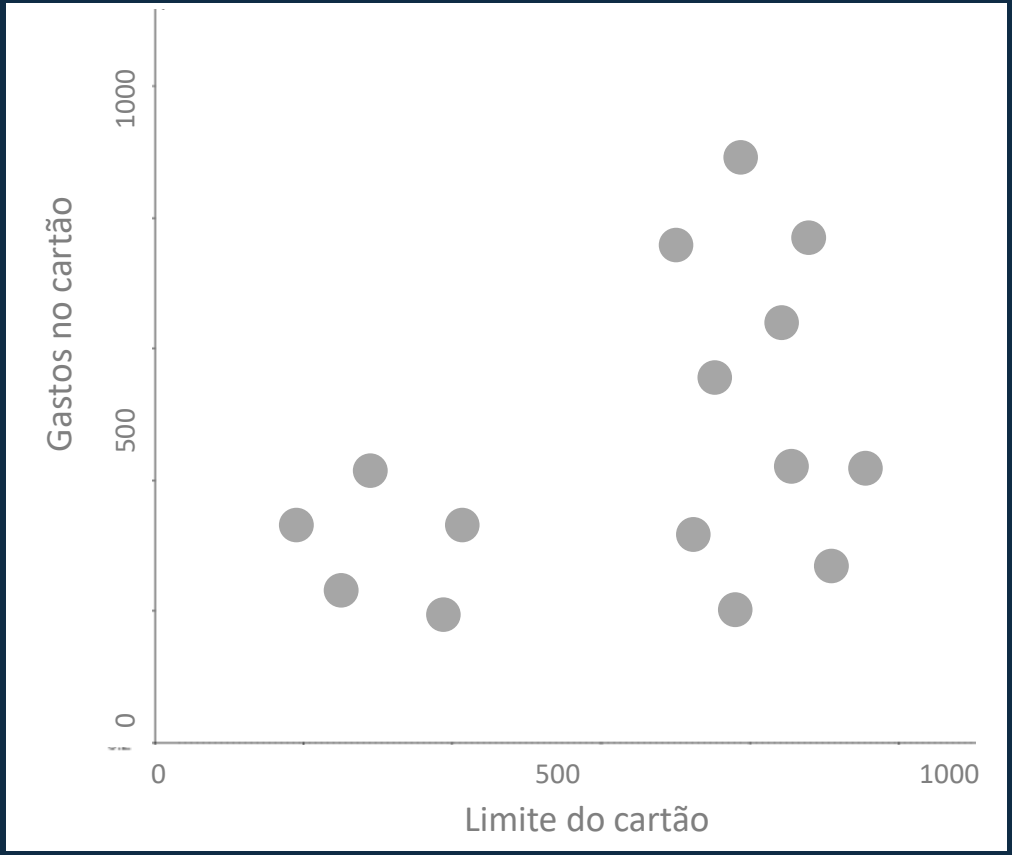


K-Means

- Um algoritmo não supervisionado
- Usado para Agrupamento (clustering)
- Algoritmo baseado em particionamento do espaço

- Derivações: K-Medoids, K-Medians

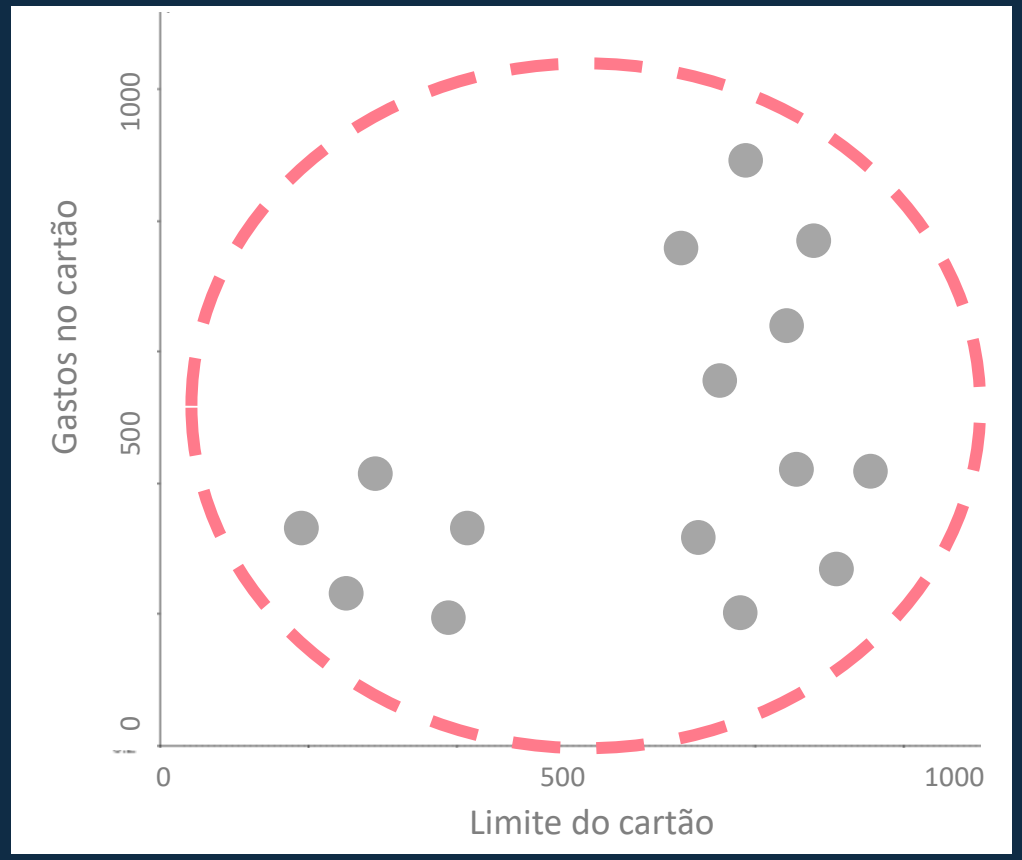
Encontrar no espaço n-dimensional
K grupos de amostras/objetos semelhantes entre si.



Intuição humana:

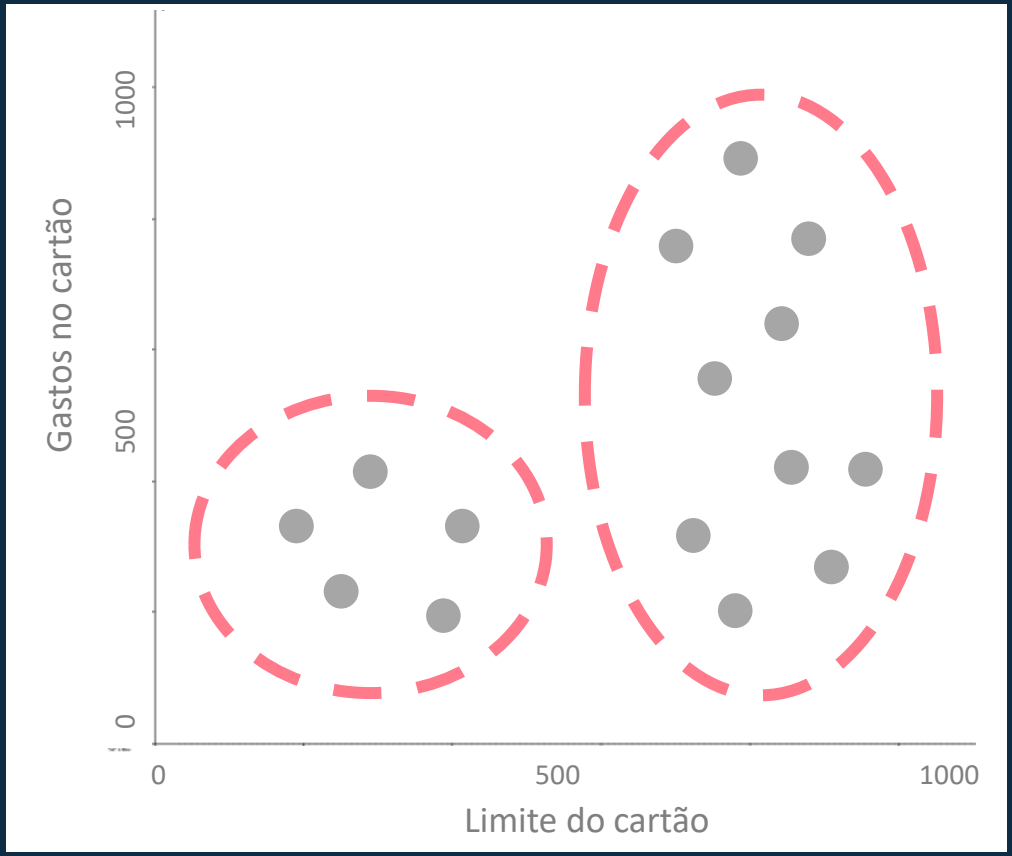
Se fosse necessário separar essas pessoas em 1, 2 e 3 grupos, como nós separaríamos?

Encontrar no espaço n-dimensional
K grupos de amostras/objetos semelhantes entre si.



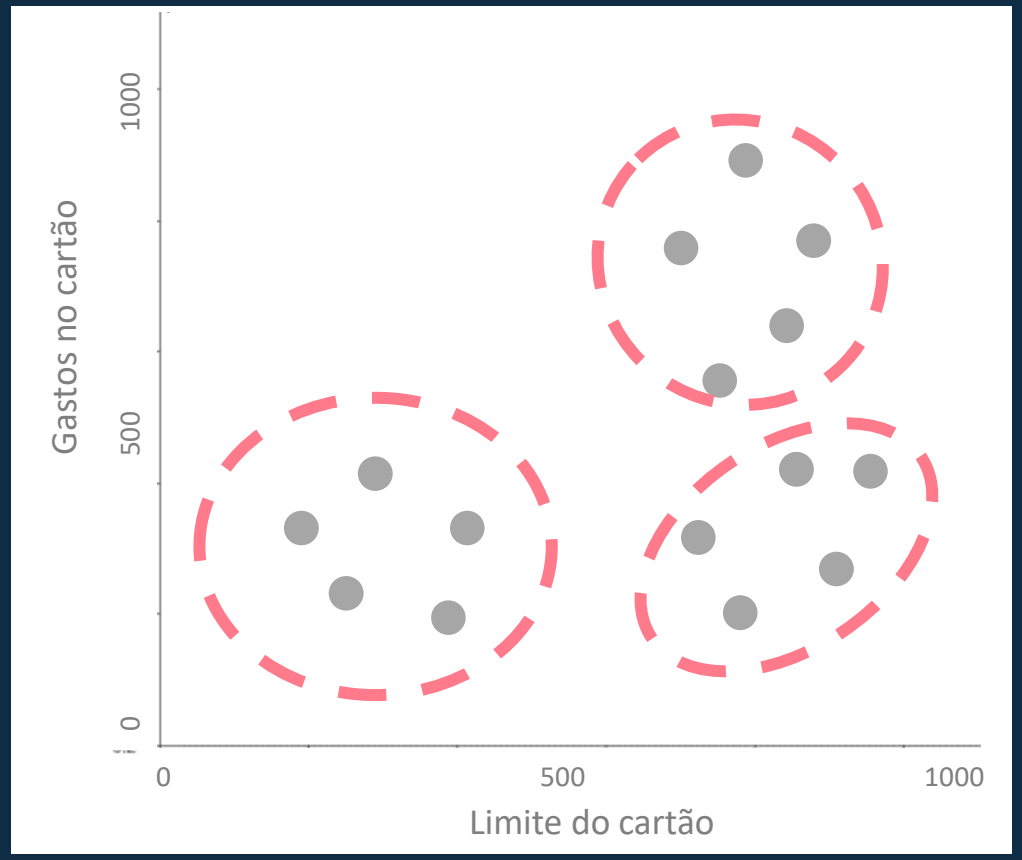
1 grupo

Encontrar no espaço n-dimensional
K grupos de amostras/objetos semelhantes entre si.



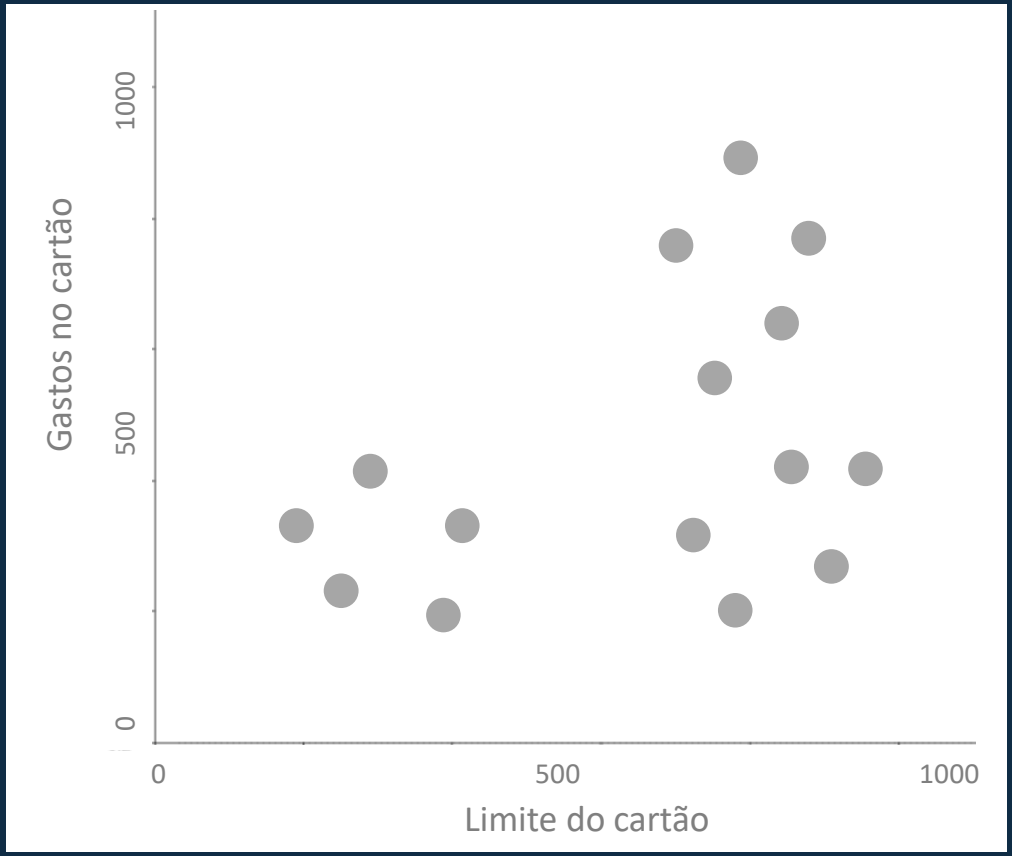
2 grupos

Encontrar no espaço n-dimensional
K grupos de amostras/objetos semelhantes entre si.



3 grupos

Steps de funcionamento do K-Means



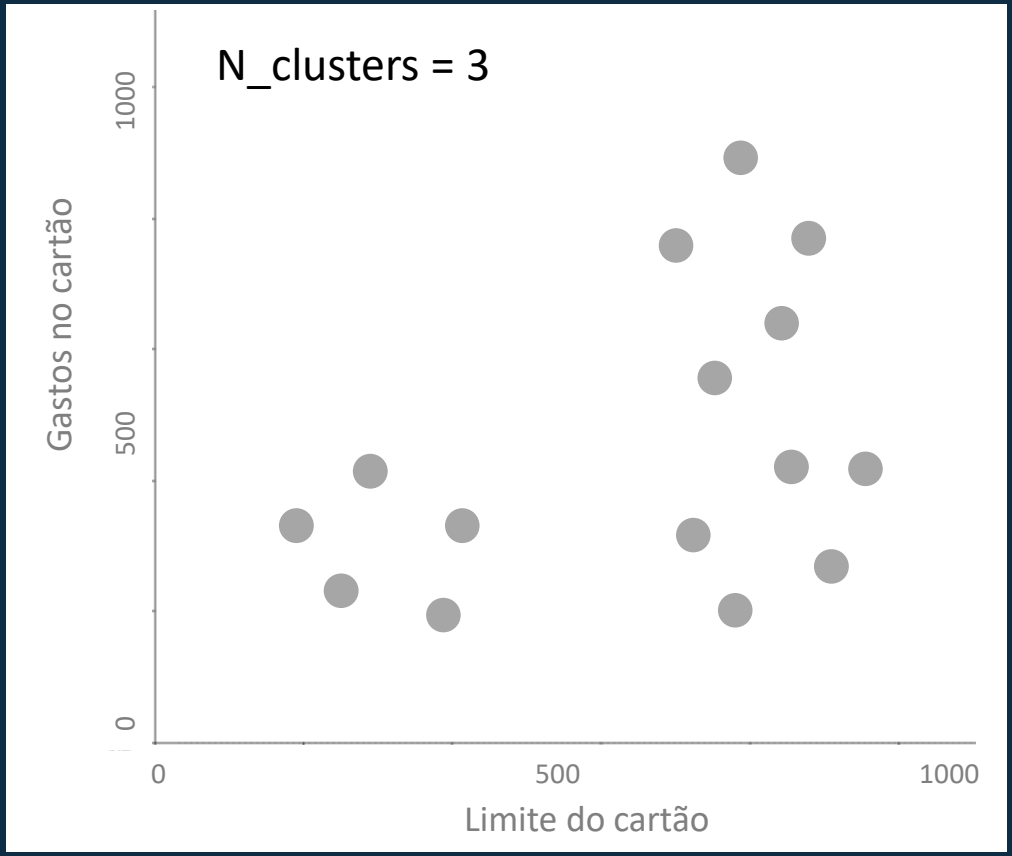
1 – Define-se o número de Clusters (K)
2 – Aleatoriamente, define-se a posição de cada centróide (centro do cluster)
3 – Calcula-se euclidiana entre cada ponto e os centróides, definindo qual é o centóide mais próximo
4 – Gera-se um novo centróide a partir da média de todos os pontos definidos em cada cluster
5 – Repete os passos 3 e 4 até que não haja mais movimentação de pontos entre clusters.

Distancia euclidiana entre A e B: $\sqrt{(x_A - xB)^2 + (yA - yB)^2}$

Ex: Ponto A (200,250) – Ponto B (320,330)

$$\sqrt{(200 - 320)^2 + (250 - 330)^2} = \mathbf{144,22}$$

Steps de funcionamento do K-Means



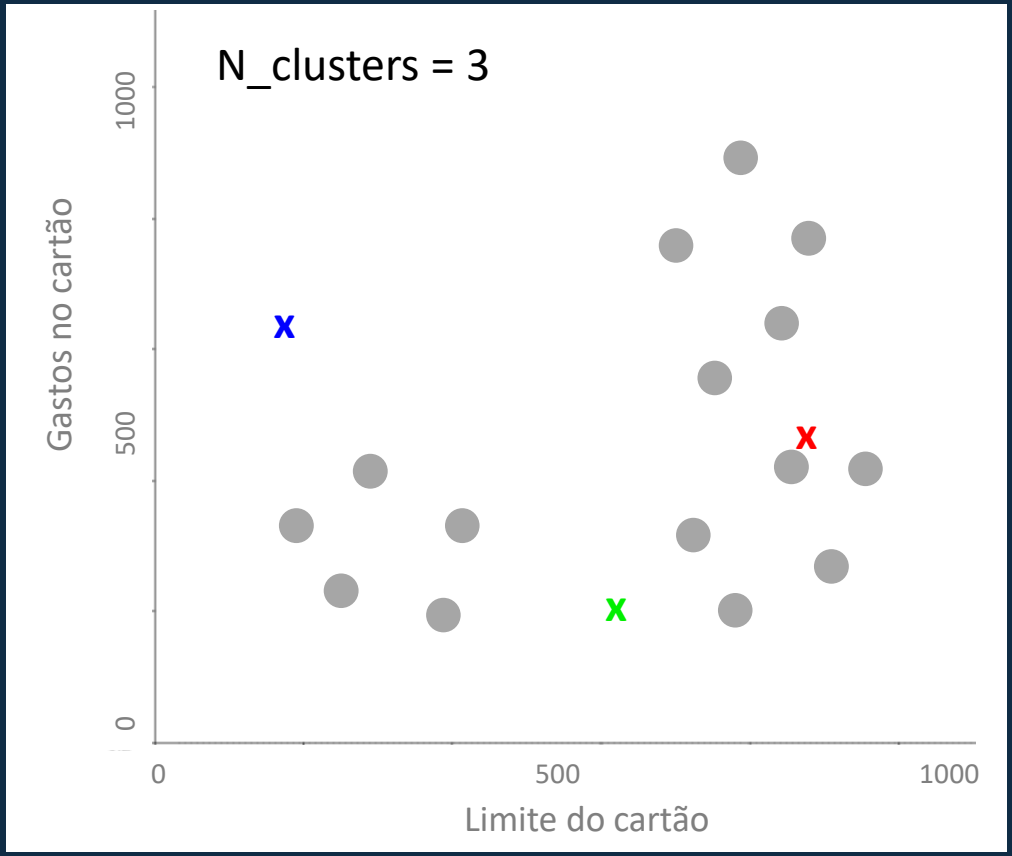
1 – Define-se o número de Clusters (K)
2 – Aleatoriamente, define-se a posição de cada centróide (centro do cluster)
3 – Calcula-se euclidiana entre cada ponto e os centróides, definindo qual é o centóide mais próximo
4 – Gera-se um novo centróide a partir da média de todos os pontos definidos em cada cluster
5 – Repete os passos 3 e 4 até que não haja mais movimentação de pontos entre clusters.

Distancia euclidiana entre A e B: $\sqrt{(x_A - xB)^2 + (yA - yB)^2}$

Ex: Ponto A (200,250) – Ponto B (320,330)

$$\sqrt{(200 - 320)^2 + (250 - 330)^2} = \mathbf{144,22}$$

Steps de funcionamento do K-Means



1 – Define-se o número de Clusters (K)

2 – Aleatoriamente, define-se a posição de cada centróide (centro do cluster)

3 – Calcula-se euclidiana entre cada ponto e os centróides, definindo qual é o centóide mais próximo

4 – Gera-se um novo centróide a partir da média de todos os pontos definidos em cada cluster

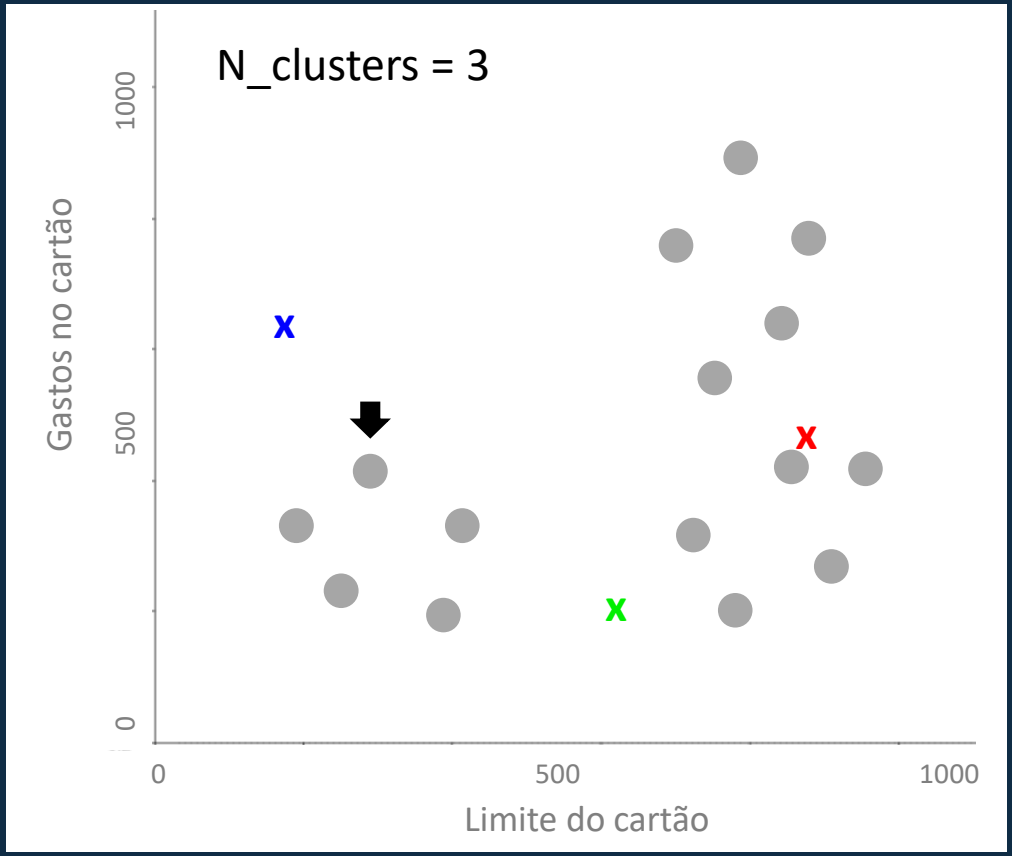
5 – Repete os passos 3 e 4 até que não haja mais movimentação de pontos entre clusters.

Distancia euclidiana entre A e B: $\sqrt{(x_A - xB)^2 + (yA - yB)^2}$

Ex: Ponto A (200,250) – Ponto B (320,330)

$$\sqrt{(200 - 320)^2 + (250 - 330)^2} = \mathbf{144,22}$$

Steps de funcionamento do K-Means



1 – Define-se o número de Clusters (K)

2 – Aleatoriamente, define-se a posição de cada centróide (centro do cluster)

3 – Calcula-se euclidiana entre cada ponto e os centróides, definindo qual é o centóide mais próximo

4 – Gera-se um novo centróide a partir da média de todos os pontos definidos em cada cluster

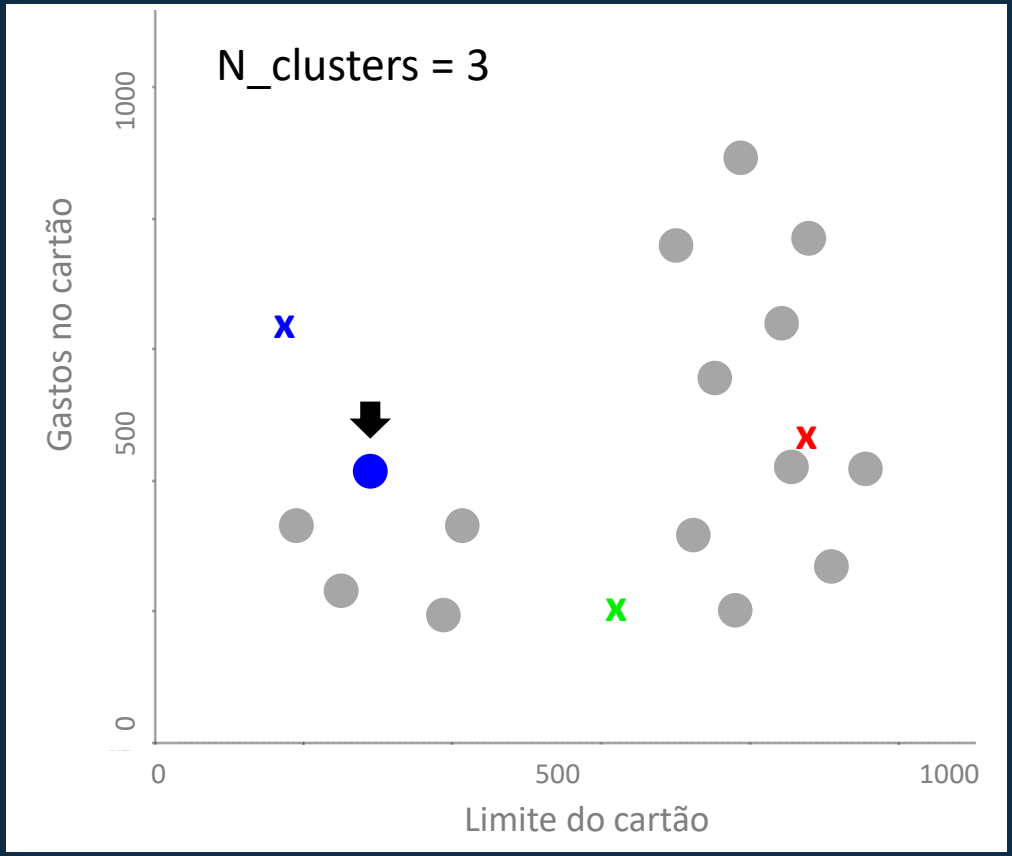
5 – Repete os passos 3 e 4 até que não haja mais movimentação de pontos entre clusters.

Distancia euclidiana entre A e B: $\sqrt{(x_A - xB)^2 + (yA - yB)^2}$

Ex: Ponto A (200,250) – Ponto B (320,330)

$$\sqrt{(200 - 320)^2 + (250 - 330)^2} = \mathbf{144,22}$$

Steps de funcionamento do K-Means



1 – Define-se o número de Clusters (K)

2 – Aleatoriamente, define-se a posição de cada centróide (centro do cluster)

3 – Calcula-se euclidiana entre cada ponto e os centróides, definindo qual é o centóide mais próximo

4 – Gera-se um novo centróide a partir da média de todos os pontos definidos em cada cluster

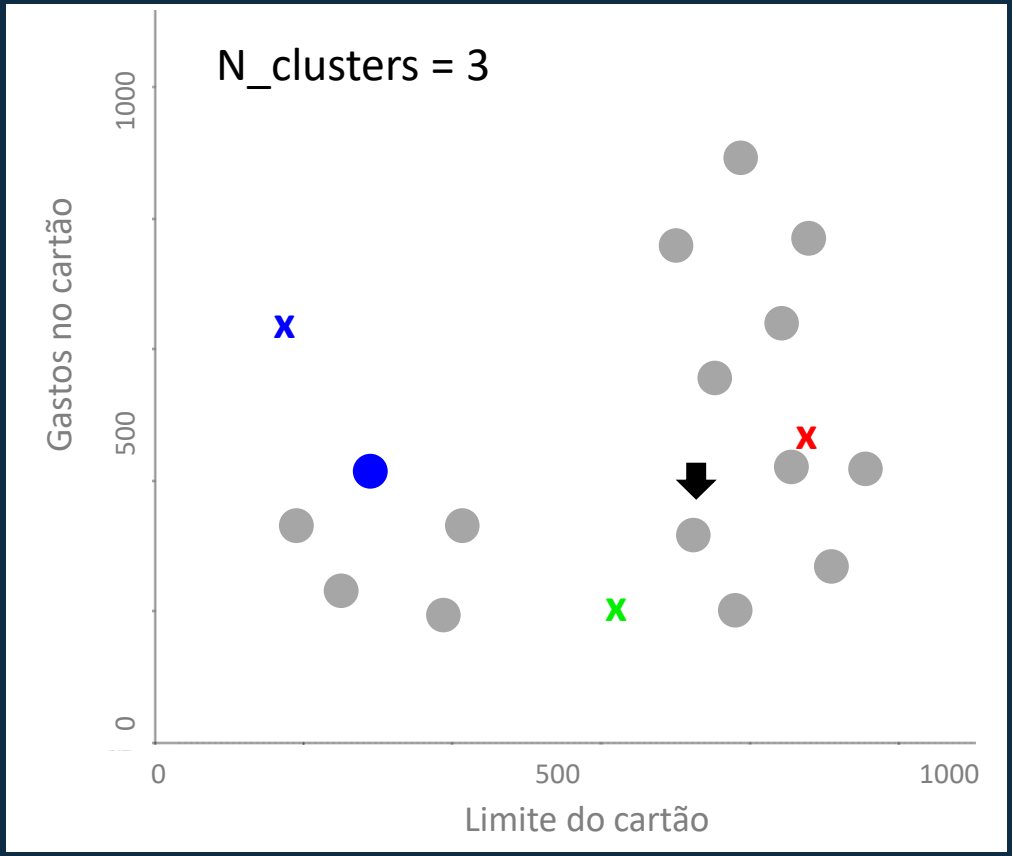
5 – Repete os passos 3 e 4 até que não haja mais movimentação de pontos entre clusters.

Distancia euclidiana entre A e B: $\sqrt{(x_A - xB)^2 + (yA - yB)^2}$

Ex: Ponto A (200,250) – Ponto B (320,330)

$$\sqrt{(200 - 320)^2 + (250 - 330)^2} = \mathbf{144,22}$$

Steps de funcionamento do K-Means



1 – Define-se o número de Clusters (K)

2 – Aleatoriamente, define-se a posição de cada centróide (centro do cluster)

3 – Calcula-se euclidiana entre cada ponto e os centróides, definindo qual é o centóide mais próximo

4 – Gera-se um novo centróide a partir da média de todos os pontos definidos em cada cluster

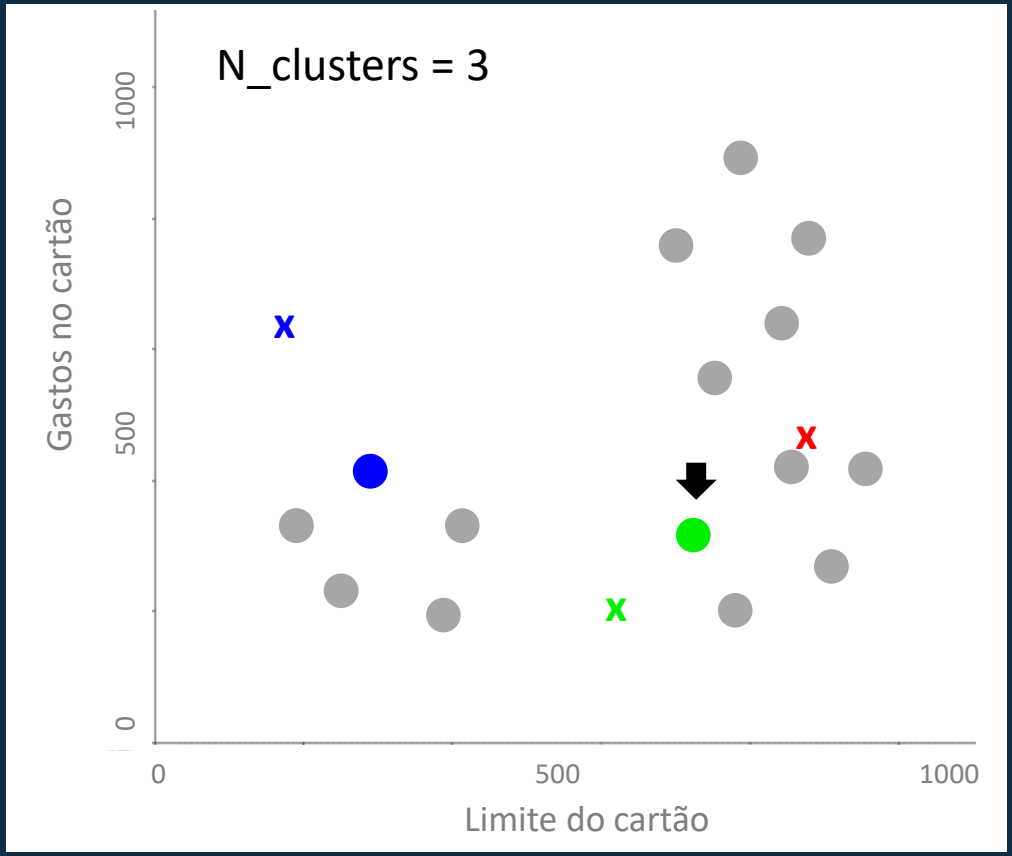
5 – Repete os passos 3 e 4 até que não haja mais movimentação de pontos entre clusters.

Distancia euclidiana entre A e B: $\sqrt{(x_A - xB)^2 + (yA - yB)^2}$

Ex: Ponto A (200,250) – Ponto B (320,330)

$$\sqrt{(200 - 320)^2 + (250 - 330)^2} = \mathbf{144,22}$$

Steps de funcionamento do K-Means



1 – Define-se o número de Clusters (K)

2 – Aleatoriamente, define-se a posição de cada centróide (centro do cluster)

3 – Calcula-se euclidiana entre cada ponto e os centróides, definindo qual é o centóide mais próximo

4 – Gera-se um novo centróide a partir da média de todos os pontos definidos em cada cluster

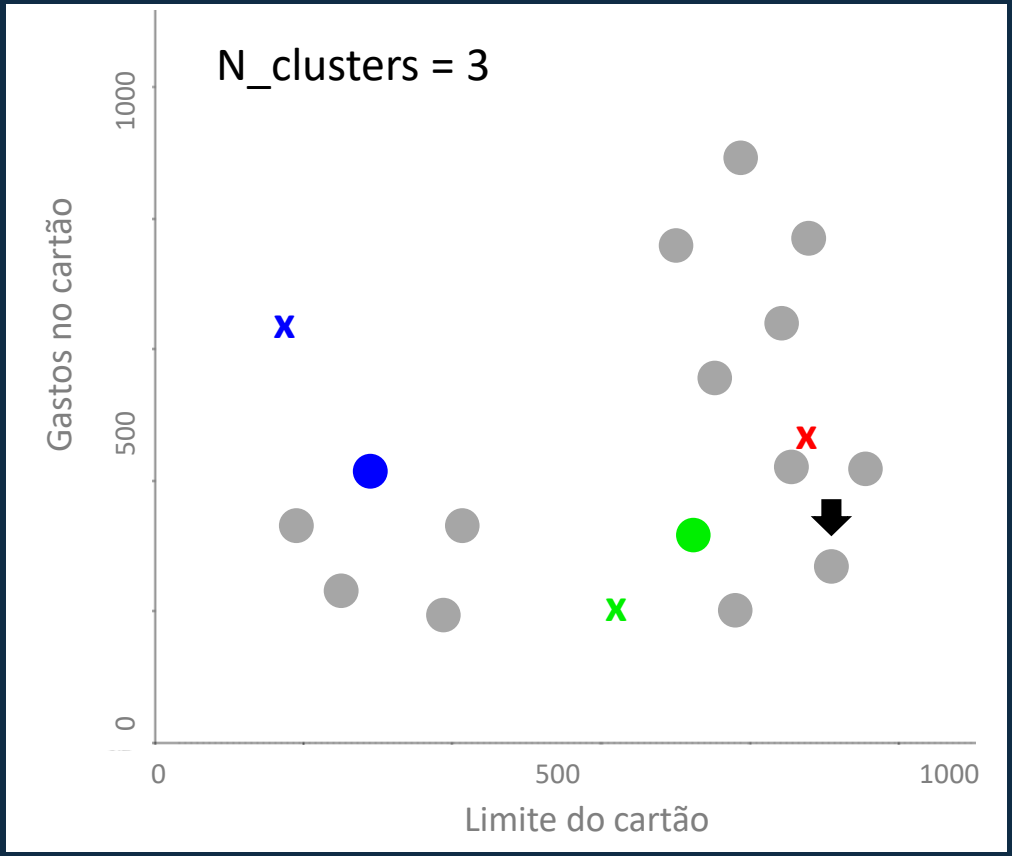
5 – Repete os passos 3 e 4 até que não haja mais movimentação de pontos entre clusters.

Distancia euclidiana entre A e B: $\sqrt{(x_A - xB)^2 + (yA - yB)^2}$

Ex: Ponto A (200,250) – Ponto B (320,330)

$$\sqrt{(200 - 320)^2 + (250 - 330)^2} = \mathbf{144,22}$$

Steps de funcionamento do K-Means



1 – Define-se o número de Clusters (K)

2 – Aleatoriamente, define-se a posição de cada centróide (centro do cluster)

3 – Calcula-se euclidiana entre cada ponto e os centróides, definindo qual é o centóide mais próximo

4 – Gera-se um novo centróide a partir da média de todos os pontos definidos em cada cluster

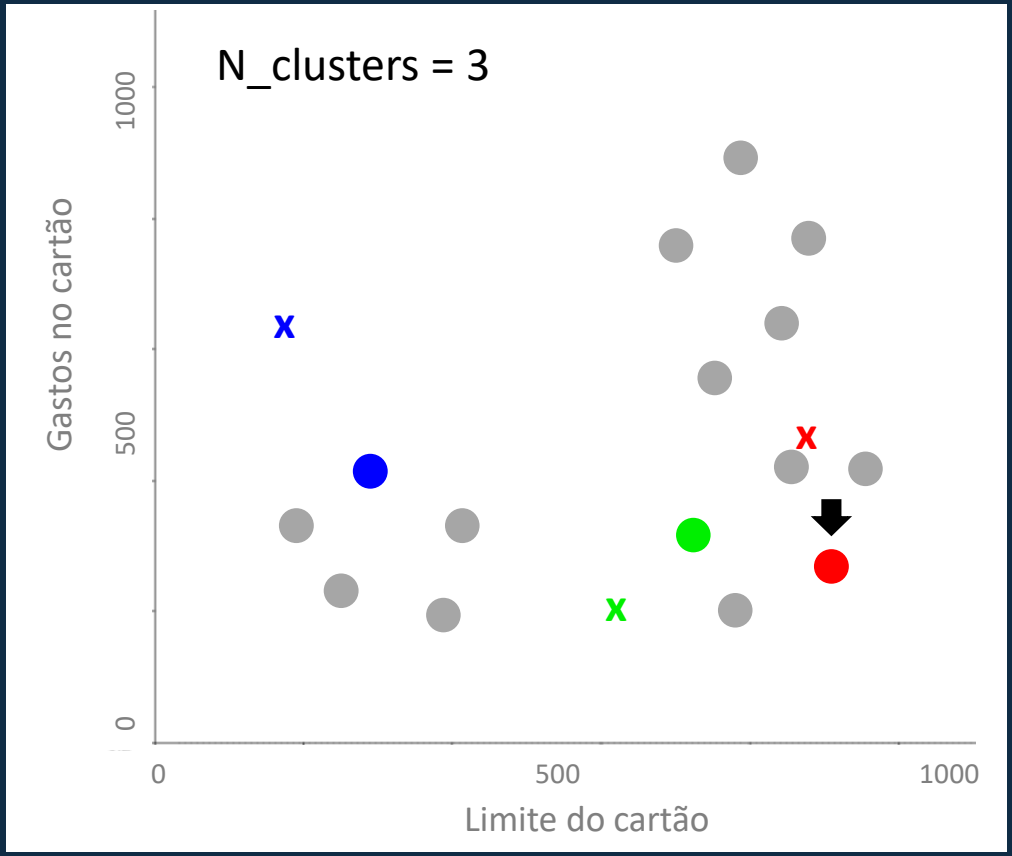
5 – Repete os passos 3 e 4 até que não haja mais movimentação de pontos entre clusters.

Distancia euclidiana entre A e B: $\sqrt{(x_A - xB)^2 + (yA - yB)^2}$

Ex: Ponto A (200,250) – Ponto B (320,330)

$$\sqrt{(200 - 320)^2 + (250 - 330)^2} = \mathbf{144,22}$$

Steps de funcionamento do K-Means



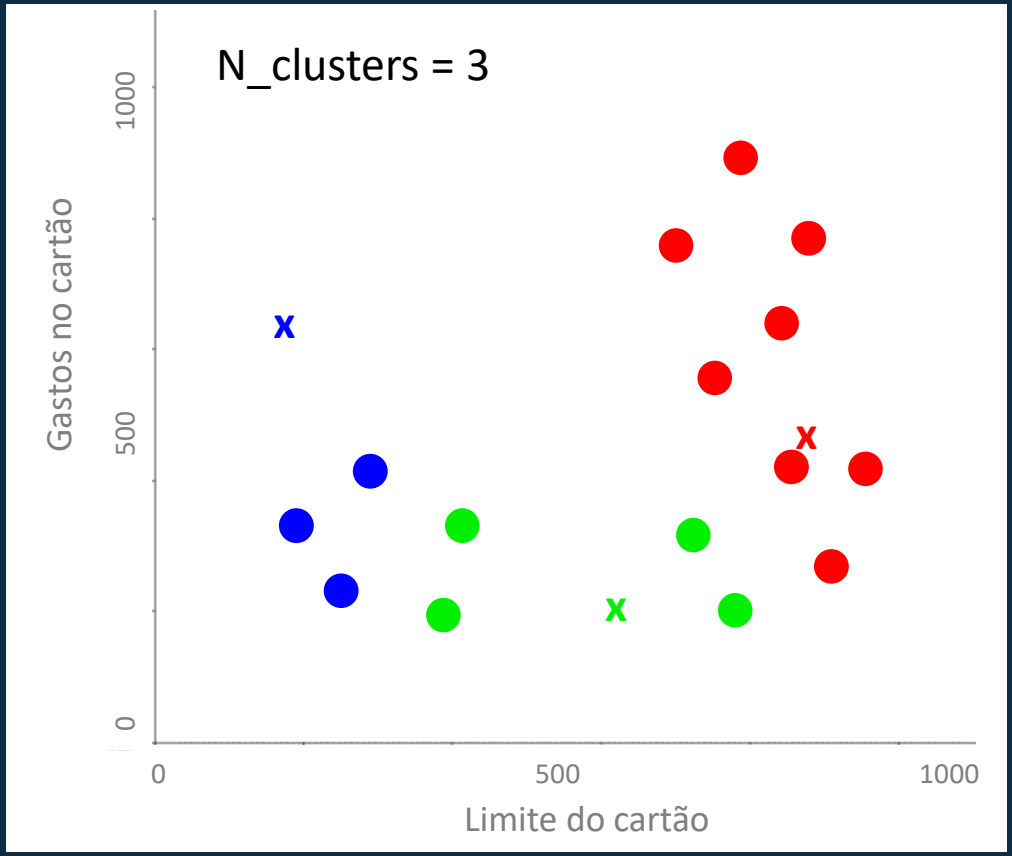
1 – Define-se o número de Clusters (K)
2 – Aleatoriamente, define-se a posição de cada centróide (centro do cluster)
3 – Calcula-se euclidiana entre cada ponto e os centróides, definindo qual é o centóide mais próximo
4 – Gera-se um novo centróide a partir da média de todos os pontos definidos em cada cluster
5 – Repete os passos 3 e 4 até que não haja mais movimentação de pontos entre clusters.

Distancia euclidiana entre A e B: $\sqrt{(x_A - xB)^2 + (yA - yB)^2}$

Ex: Ponto A (200,250) – Ponto B (320,330)

$$\sqrt{(200 - 320)^2 + (250 - 330)^2} = \mathbf{144,22}$$

Steps de funcionamento do K-Means



1 – Define-se o número de Clusters (K)

2 – Aleatoriamente, define-se a posição de cada centróide (centro do cluster)

3 – Calcula-se euclidiana entre cada ponto e os centróides, definindo qual é o centóide mais próximo

4 – Gera-se um novo centróide a partir da média de todos os pontos definidos em cada cluster

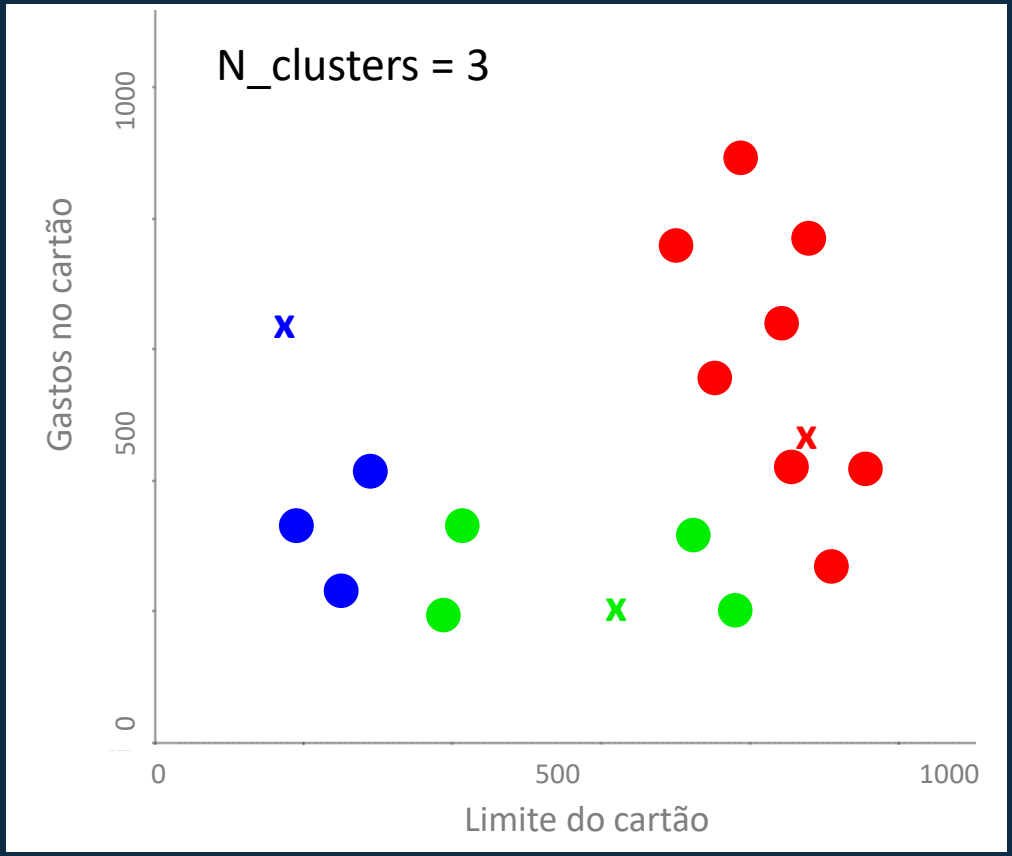
5 – Repete os passos 3 e 4 até que não haja mais movimentação de pontos entre clusters.

Distancia euclidiana entre A e B: $\sqrt{(x_A - xB)^2 + (yA - yB)^2}$

Ex: Ponto A (200,250) – Ponto B (320,330)

$$\sqrt{(200 - 320)^2 + (250 - 330)^2} = \mathbf{144,22}$$

Steps de funcionamento do K-Means



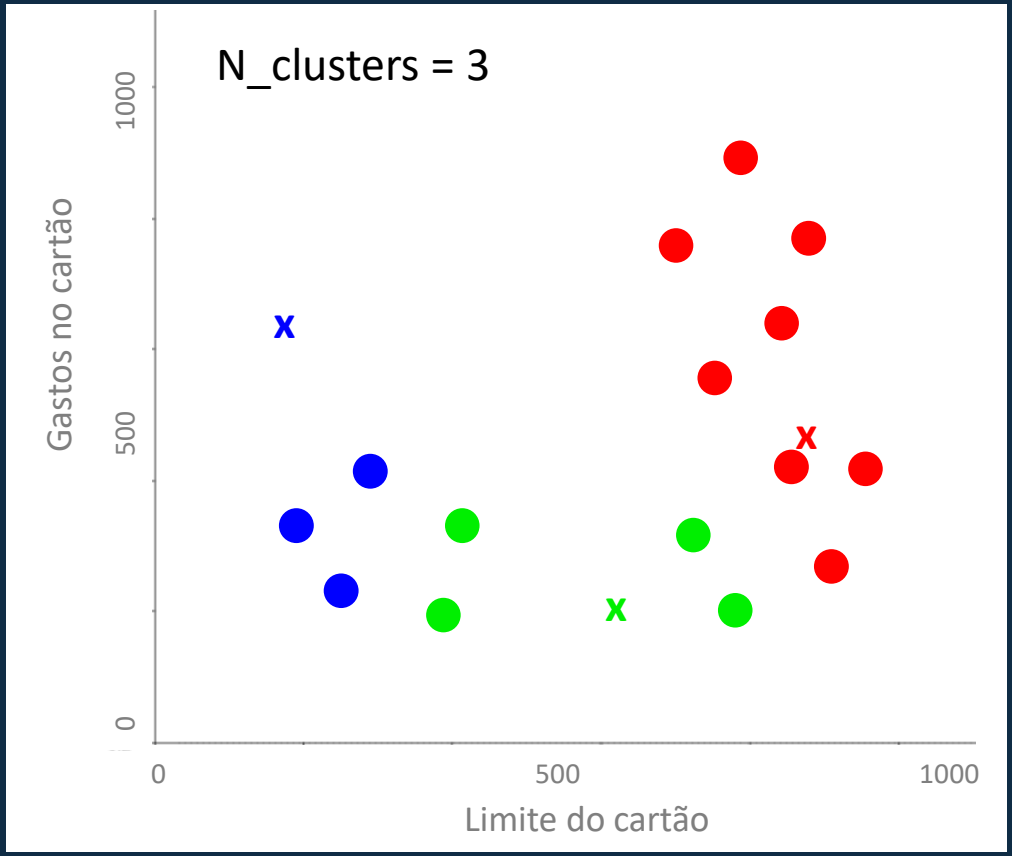
1 – Define-se o número de Clusters (K)
2 – Aleatoriamente, define-se a posição de cada centróide (centro do cluster)
3 – Calcula-se euclidiana entre cada ponto e os centróides, definindo qual é o centóide mais próximo
4 – Gera-se um novo centróide a partir da média de todos os pontos definidos em cada cluster
5 – Repete os passos 3 e 4 até que não haja mais movimentação de pontos entre clusters.

Distancia euclidiana entre A e B: $\sqrt{(x_A - xB)^2 + (yA - yB)^2}$

Ex: Ponto A (200,250) – Ponto B (320,330)

$$\sqrt{(200 - 320)^2 + (250 - 330)^2} = \mathbf{144,22}$$

Steps de funcionamento do K-Means



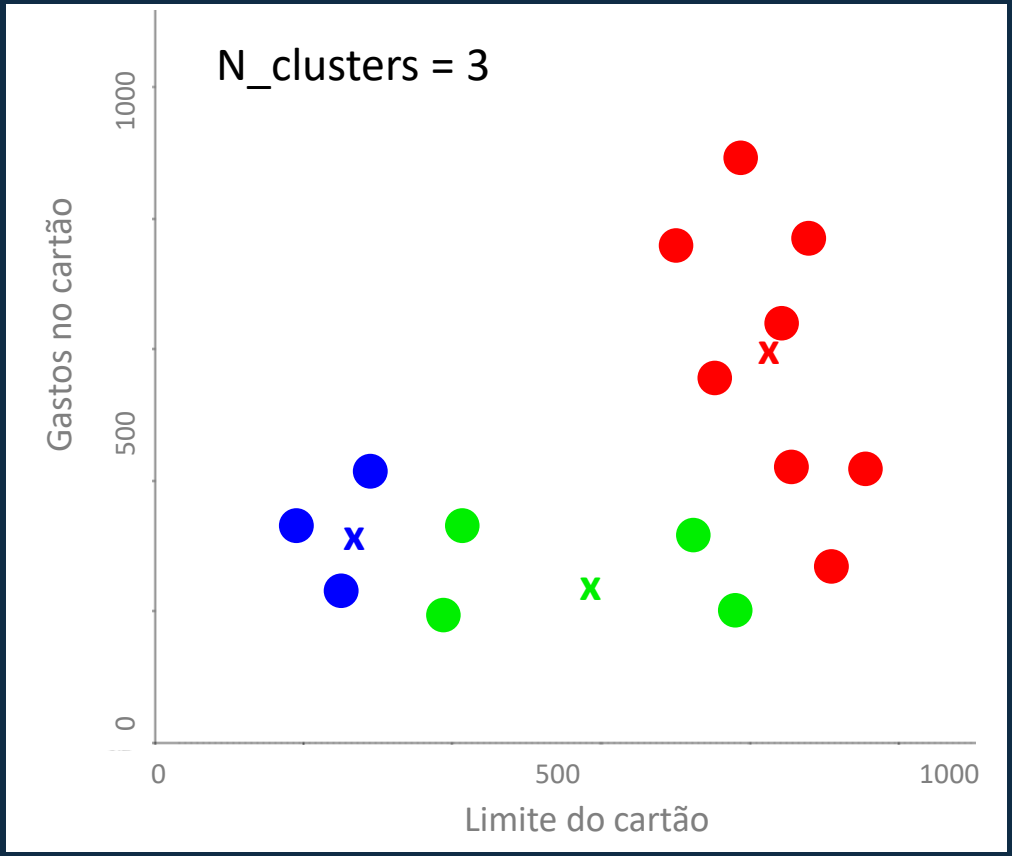
1 – Define-se o número de Clusters (K)
2 – Aleatoriamente, define-se a posição de cada centróide (centro do cluster)
3 – Calcula-se euclidiana entre cada ponto e os centróides, definindo qual é o centóide mais próximo
4 – Gera-se um novo centróide a partir da média de todos os pontos definidos em cada cluster
5 – Repete os passos 3 e 4 até que não haja mais movimentação de pontos entre clusters.

Distancia euclidiana entre A e B: $\sqrt{(x_A - xB)^2 + (yA - yB)^2}$

Ex: Ponto A (200,250) – Ponto B (320,330)

$$\sqrt{(200 - 320)^2 + (250 - 330)^2} = \mathbf{144,22}$$

Steps de funcionamento do K-Means



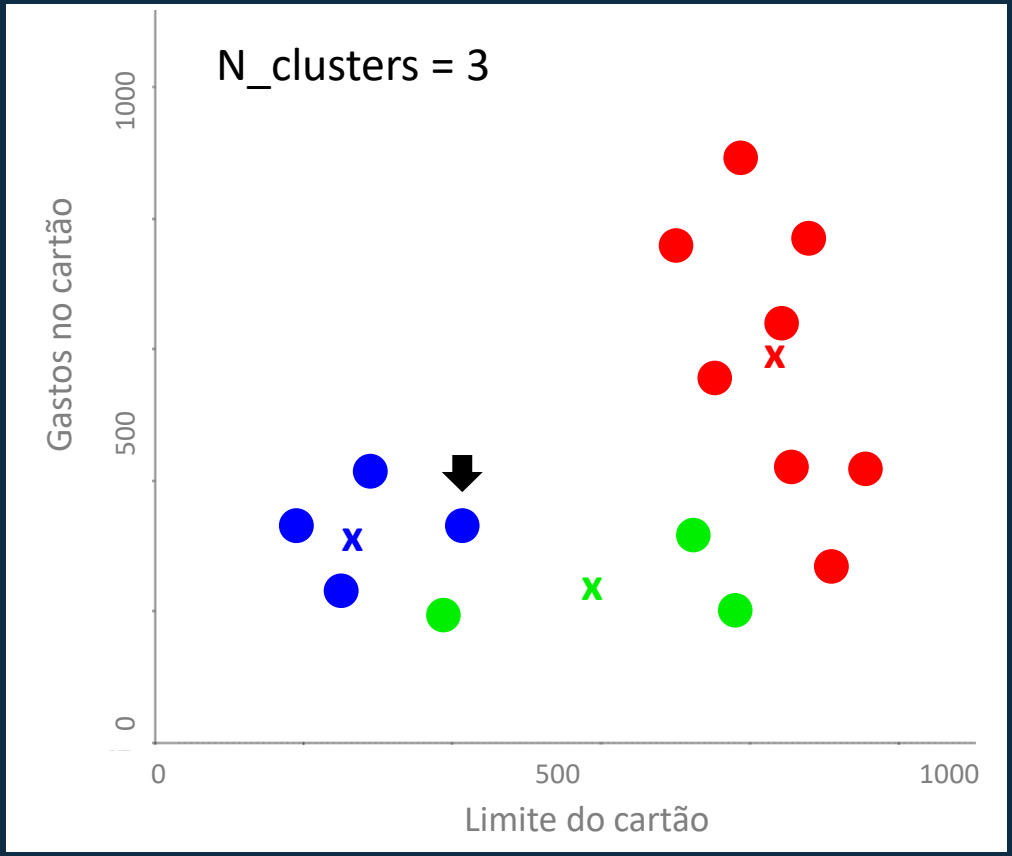
1 – Define-se o número de Clusters (K)
2 – Aleatoriamente, define-se a posição de cada centróide (centro do cluster)
3 – Calcula-se euclidiana entre cada ponto e os centróides, definindo qual é o centóide mais próximo
4 – Gera-se um novo centróide a partir da média de todos os pontos definidos em cada cluster
5 – Repete os passos 3 e 4 até que não haja mais movimentação de pontos entre clusters.

Distancia euclidiana entre A e B: $\sqrt{(x_A - xB)^2 + (yA - yB)^2}$

Ex: Ponto A (200,250) – Ponto B (320,330)

$$\sqrt{(200 - 320)^2 + (250 - 330)^2} = \mathbf{144,22}$$

Steps de funcionamento do K-Means



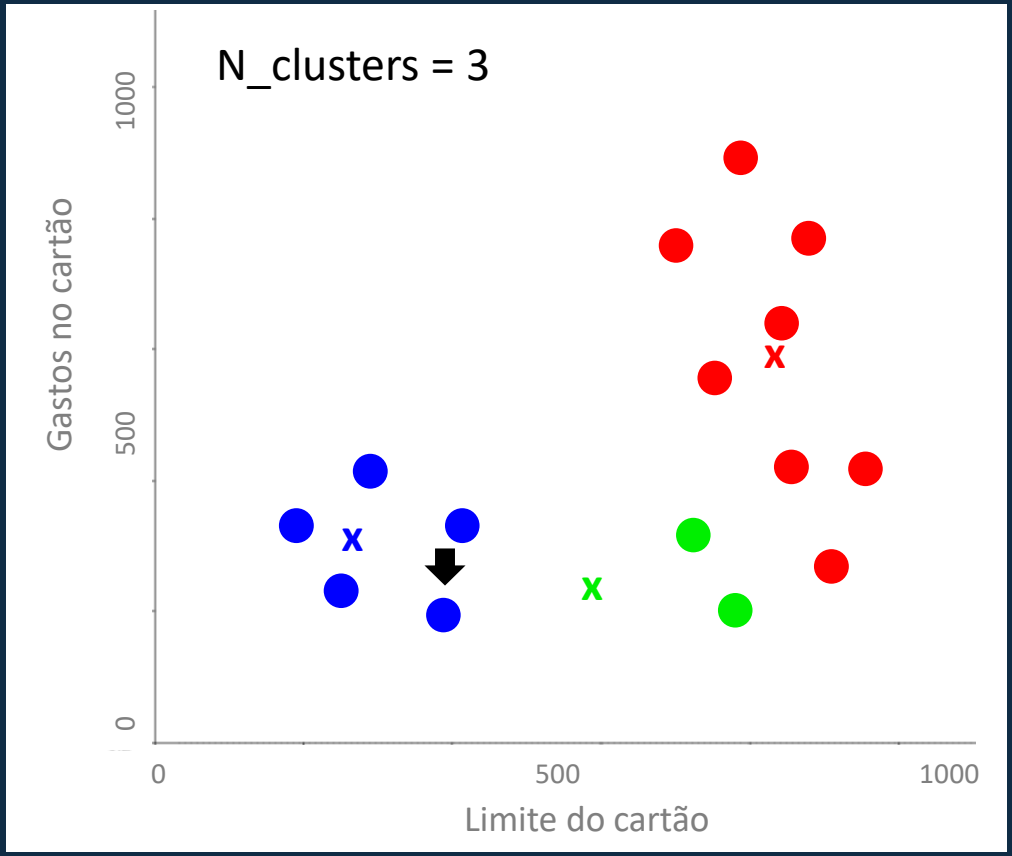
1 – Define-se o número de Clusters (K)
2 – Aleatoriamente, define-se a posição de cada centróide (centro do cluster)
3 – Calcula-se euclidiana entre cada ponto e os centróides, definindo qual é o centóide mais próximo
4 – Gera-se um novo centróide a partir da média de todos os pontos definidos em cada cluster
5 – Repete os passos 3 e 4 até que não haja mais movimentação de pontos entre clusters.

Distancia euclidiana entre A e B: $\sqrt{(x_A - xB)^2 + (yA - yB)^2}$

Ex: Ponto A (200,250) – Ponto B (320,330)

$$\sqrt{(200 - 320)^2 + (250 - 330)^2} = \mathbf{144,22}$$

Steps de funcionamento do K-Means



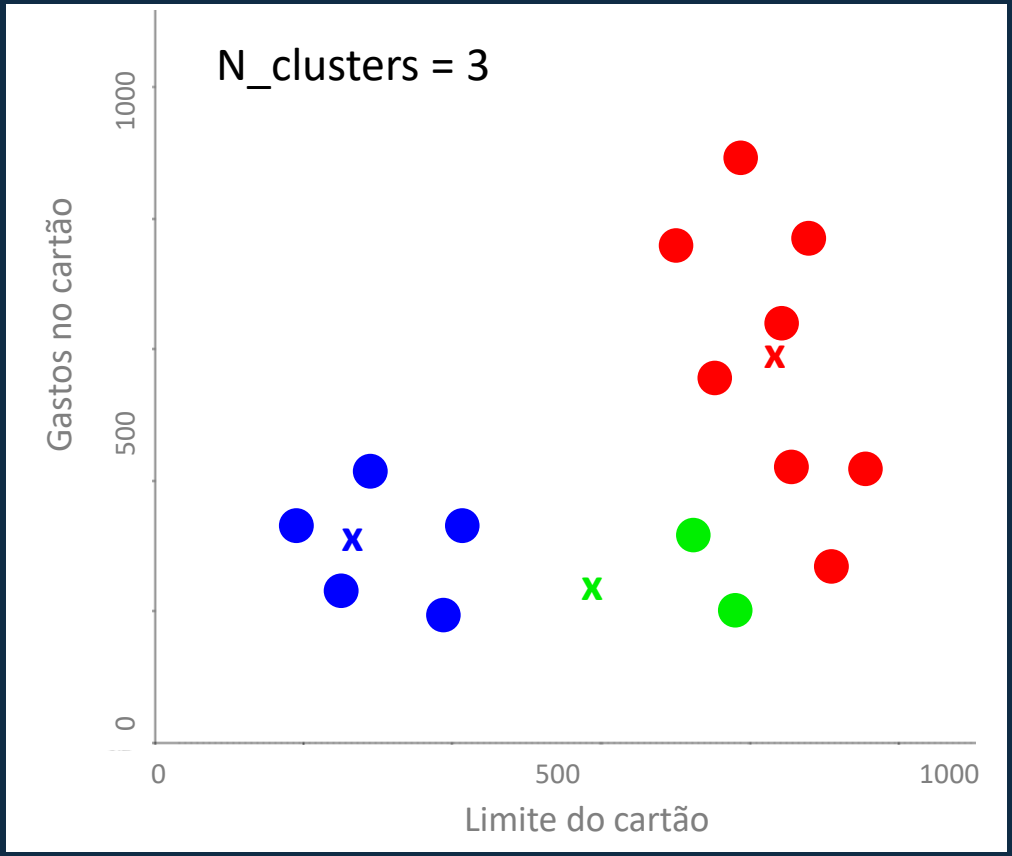
1 – Define-se o número de Clusters (K)
2 – Aleatoriamente, define-se a posição de cada centróide (centro do cluster)
3 – Calcula-se euclidiana entre cada ponto e os centróides, definindo qual é o centóide mais próximo
4 – Gera-se um novo centróide a partir da média de todos os pontos definidos em cada cluster
5 – Repete os passos 3 e 4 até que não haja mais movimentação de pontos entre clusters.

Distancia euclidiana entre A e B: $\sqrt{(x_A - xB)^2 + (yA - yB)^2}$

Ex: Ponto A (200,250) – Ponto B (320,330)

$$\sqrt{(200 - 320)^2 + (250 - 330)^2} = \mathbf{144,22}$$

Steps de funcionamento do K-Means



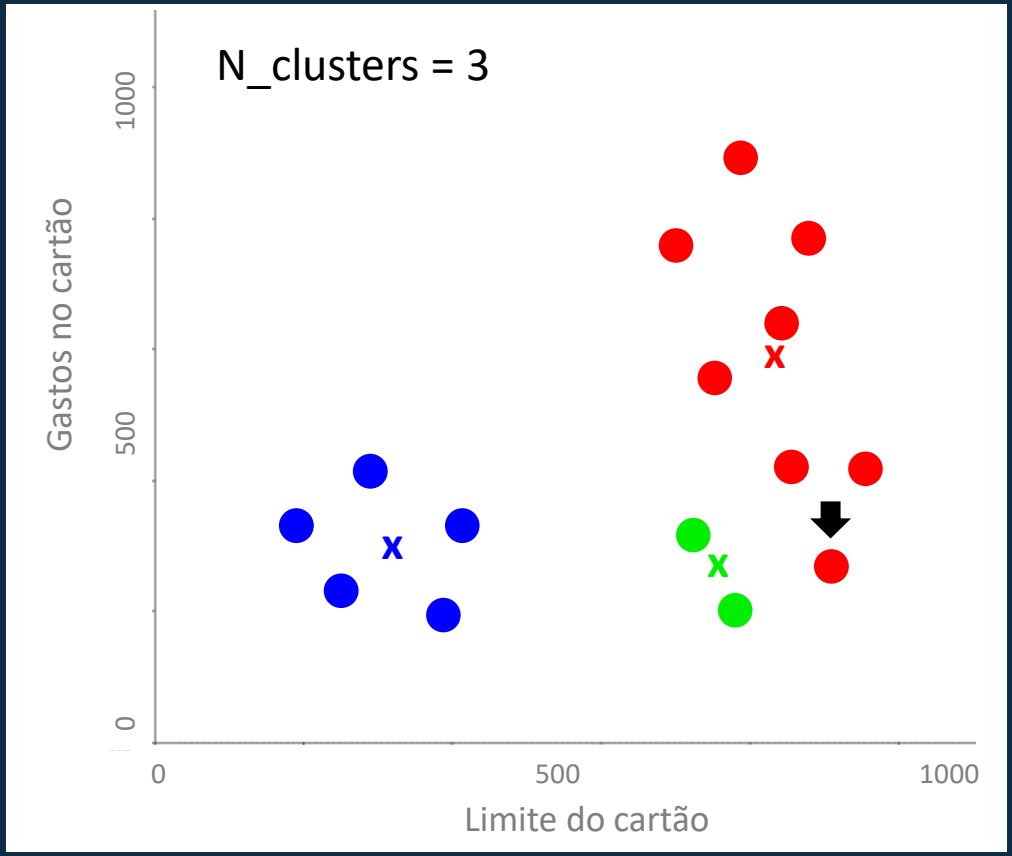
1 – Define-se o número de Clusters (K)
2 – Aleatoriamente, define-se a posição de cada centróide (centro do cluster)
3 – Calcula-se euclidiana entre cada ponto e os centróides, definindo qual é o centóide mais próximo
4 – Gera-se um novo centróide a partir da média de todos os pontos definidos em cada cluster
5 – Repete os passos 3 e 4 até que não haja mais movimentação de pontos entre clusters.

Distancia euclidiana entre A e B: $\sqrt{(x_A - xB)^2 + (yA - yB)^2}$

Ex: Ponto A (200,250) – Ponto B (320,330)

$$\sqrt{(200 - 320)^2 + (250 - 330)^2} = \mathbf{144,22}$$

Steps de funcionamento do K-Means



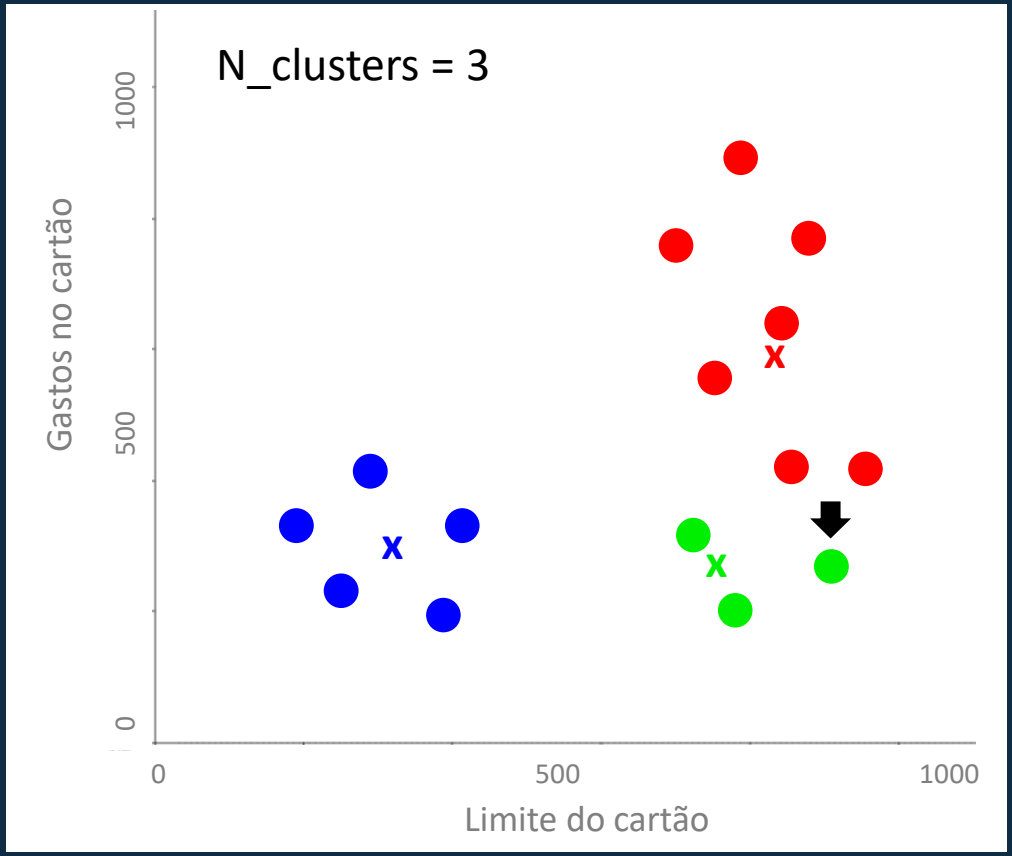
1 – Define-se o número de Clusters (K)
2 – Aleatoriamente, define-se a posição de cada centróide (centro do cluster)
3 – Calcula-se euclidiana entre cada ponto e os centróides, definindo qual é o centóide mais próximo
4 – Gera-se um novo centróide a partir da média de todos os pontos definidos em cada cluster
5 – Repete os passos 3 e 4 até que não haja mais movimentação de pontos entre clusters.

Distancia euclidiana entre A e B: $\sqrt{(x_A - xB)^2 + (yA - yB)^2}$

Ex: Ponto A (200,250) – Ponto B (320,330)

$$\sqrt{(200 - 320)^2 + (250 - 330)^2} = \mathbf{144,22}$$

Steps de funcionamento do K-Means



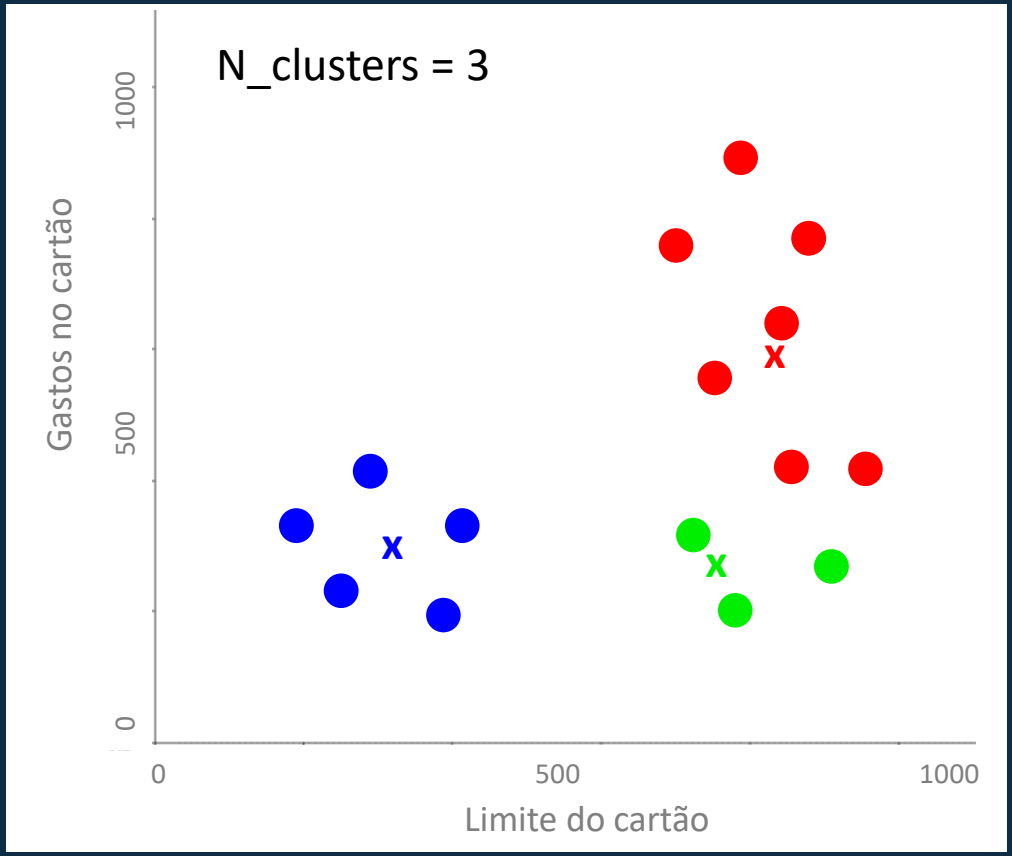
1 – Define-se o número de Clusters (K)
2 – Aleatoriamente, define-se a posição de cada centróide (centro do cluster)
3 – Calcula-se euclidiana entre cada ponto e os centróides, definindo qual é o centóide mais próximo
4 – Gera-se um novo centróide a partir da média de todos os pontos definidos em cada cluster
5 – Repete os passos 3 e 4 até que não haja mais movimentação de pontos entre clusters.

Distancia euclidiana entre A e B: $\sqrt{(x_A - xB)^2 + (yA - yB)^2}$

Ex: Ponto A (200,250) – Ponto B (320,330)

$$\sqrt{(200 - 320)^2 + (250 - 330)^2} = \mathbf{144,22}$$

Steps de funcionamento do K-Means



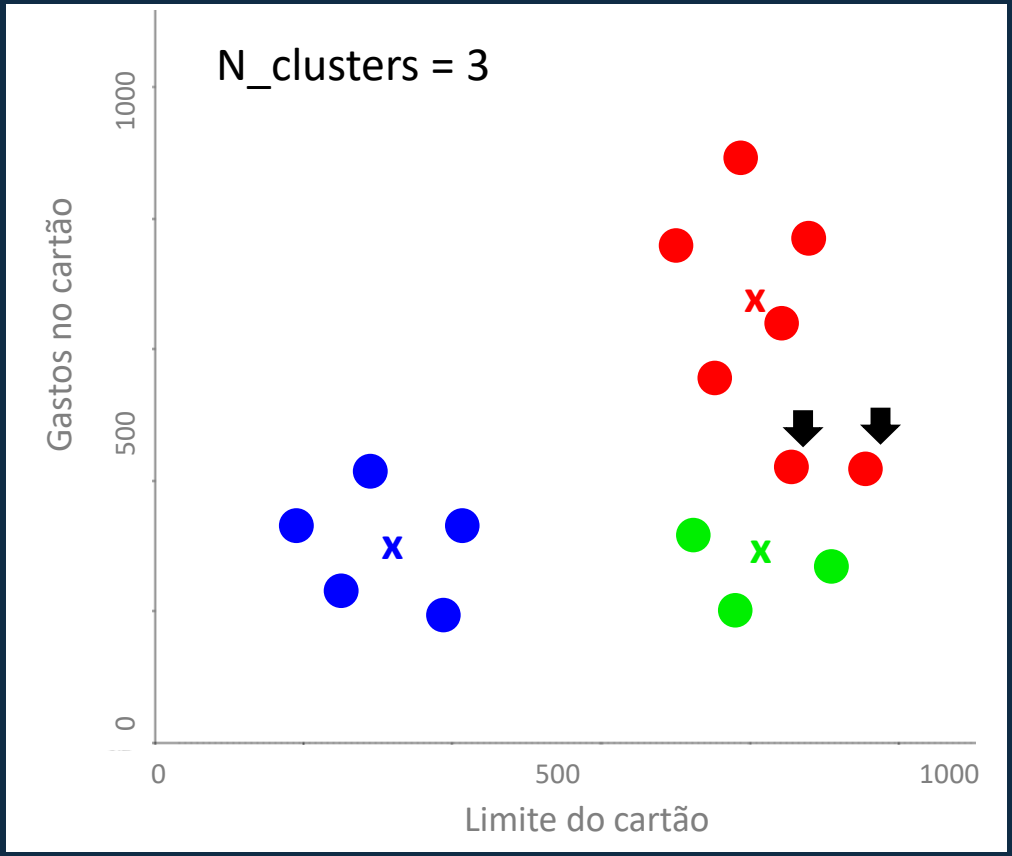
1 – Define-se o número de Clusters (K)
2 – Aleatoriamente, define-se a posição de cada centróide (centro do cluster)
3 – Calcula-se euclidiana entre cada ponto e os centróides, definindo qual é o centóide mais próximo
4 – Gera-se um novo centróide a partir da média de todos os pontos definidos em cada cluster
5 – Repete os passos 3 e 4 até que não haja mais movimentação de pontos entre clusters.

Distancia euclidiana entre A e B: $\sqrt{(x_A - xB)^2 + (yA - yB)^2}$

Ex: Ponto A (200,250) – Ponto B (320,330)

$$\sqrt{(200 - 320)^2 + (250 - 330)^2} = \mathbf{144,22}$$

Steps de funcionamento do K-Means



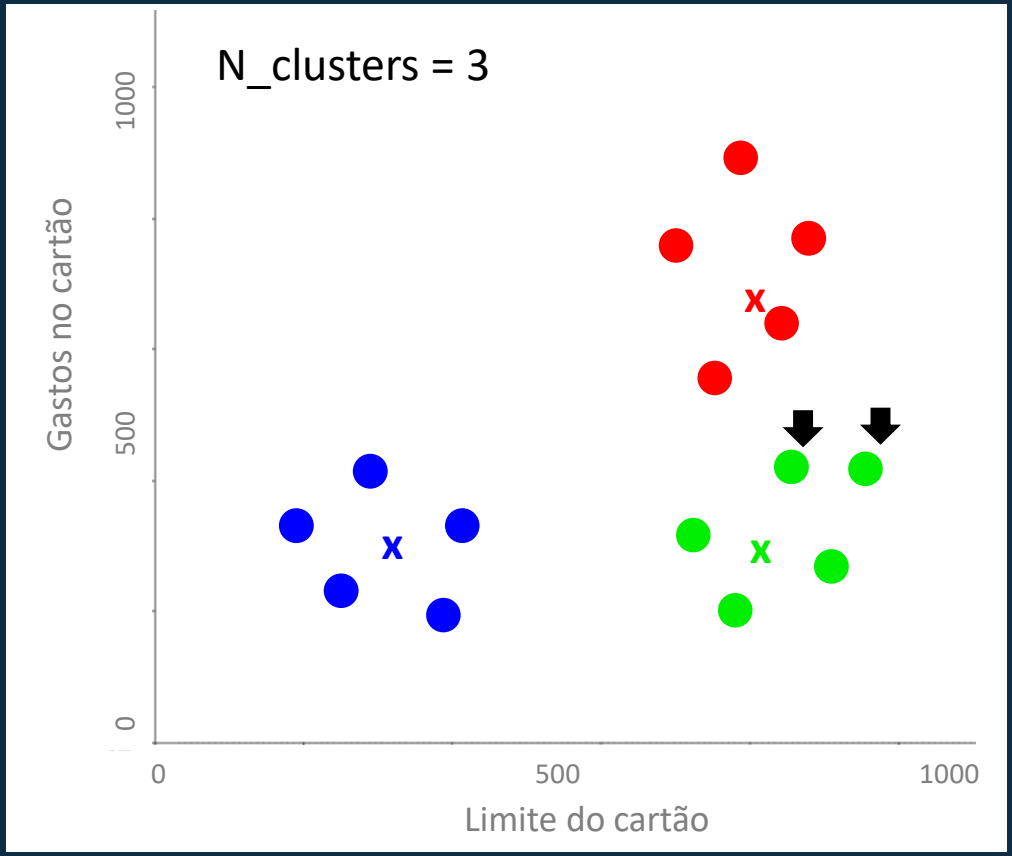
1 – Define-se o número de Clusters (K)
2 – Aleatoriamente, define-se a posição de cada centróide (centro do cluster)
3 – Calcula-se euclidiana entre cada ponto e os centróides, definindo qual é o centóide mais próximo
4 – Gera-se um novo centróide a partir da média de todos os pontos definidos em cada cluster
5 – Repete os passos 3 e 4 até que não haja mais movimentação de pontos entre clusters.

Distancia euclidiana entre A e B: $\sqrt{(x_A - xB)^2 + (yA - yB)^2}$

Ex: Ponto A (200,250) – Ponto B (320,330)

$$\sqrt{(200 - 320)^2 + (250 - 330)^2} = \mathbf{144,22}$$

Steps de funcionamento do K-Means



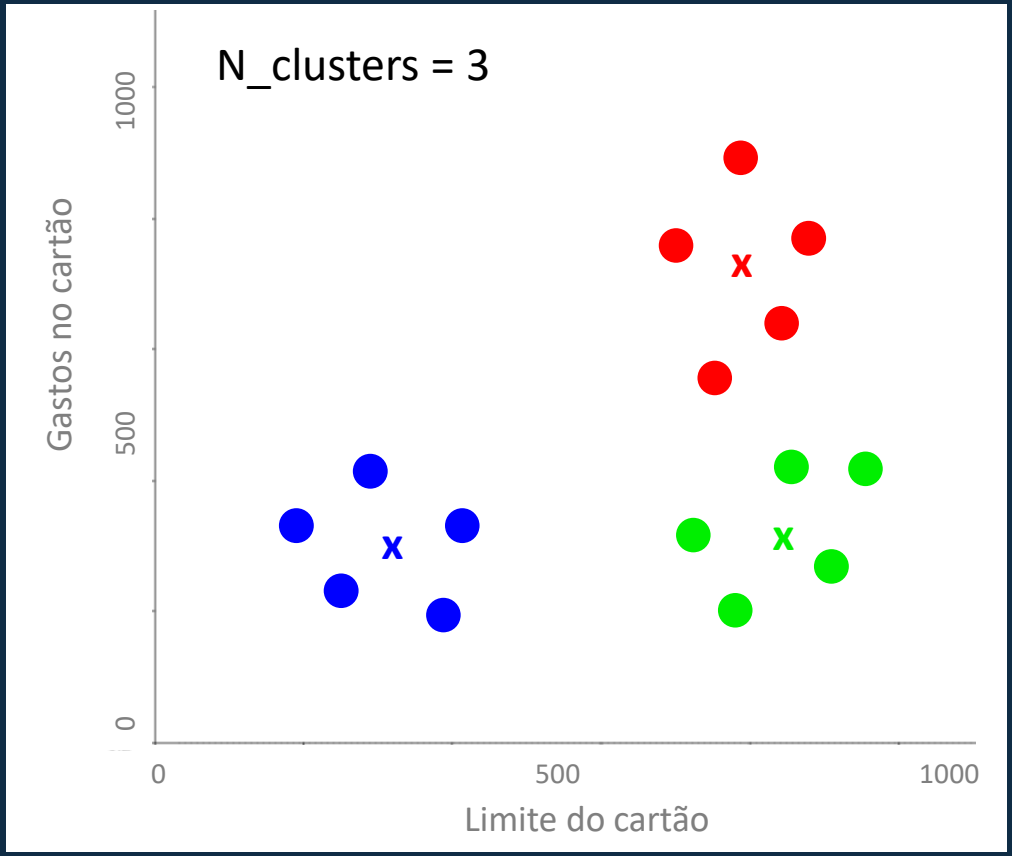
1 – Define-se o número de Clusters (K)
2 – Aleatoriamente, define-se a posição de cada centróide (centro do cluster)
3 – Calcula-se euclidiana entre cada ponto e os centróides, definindo qual é o centóide mais próximo
4 – Gera-se um novo centróide a partir da média de todos os pontos definidos em cada cluster
5 – Repete os passos 3 e 4 até que não haja mais movimentação de pontos entre clusters.

Distancia euclidiana entre A e B: $\sqrt{(x_A - xB)^2 + (yA - yB)^2}$

Ex: Ponto A (200,250) – Ponto B (320,330)

$$\sqrt{(200 - 320)^2 + (250 - 330)^2} = \mathbf{144,22}$$

Steps de funcionamento do K-Means



1 – Define-se o número de Clusters (K)
2 – Aleatoriamente, define-se a posição de cada centróide (centro do cluster)
3 – Calcula-se euclidiana entre cada ponto e os centróides, definindo qual é o centóide mais próximo
4 – Gera-se um novo centróide a partir da média de todos os pontos definidos em cada cluster
5 – Repete os passos 3 e 4 até que não haja mais movimentação de pontos entre clusters.

Distancia euclidiana entre A e B: $\sqrt{(x_A - xB)^2 + (yA - yB)^2}$

Ex: Ponto A (200,250) – Ponto B (320,330)

$$\sqrt{(200 - 320)^2 + (250 - 330)^2} = \mathbf{144,22}$$

Como definir o número correto de clusters?

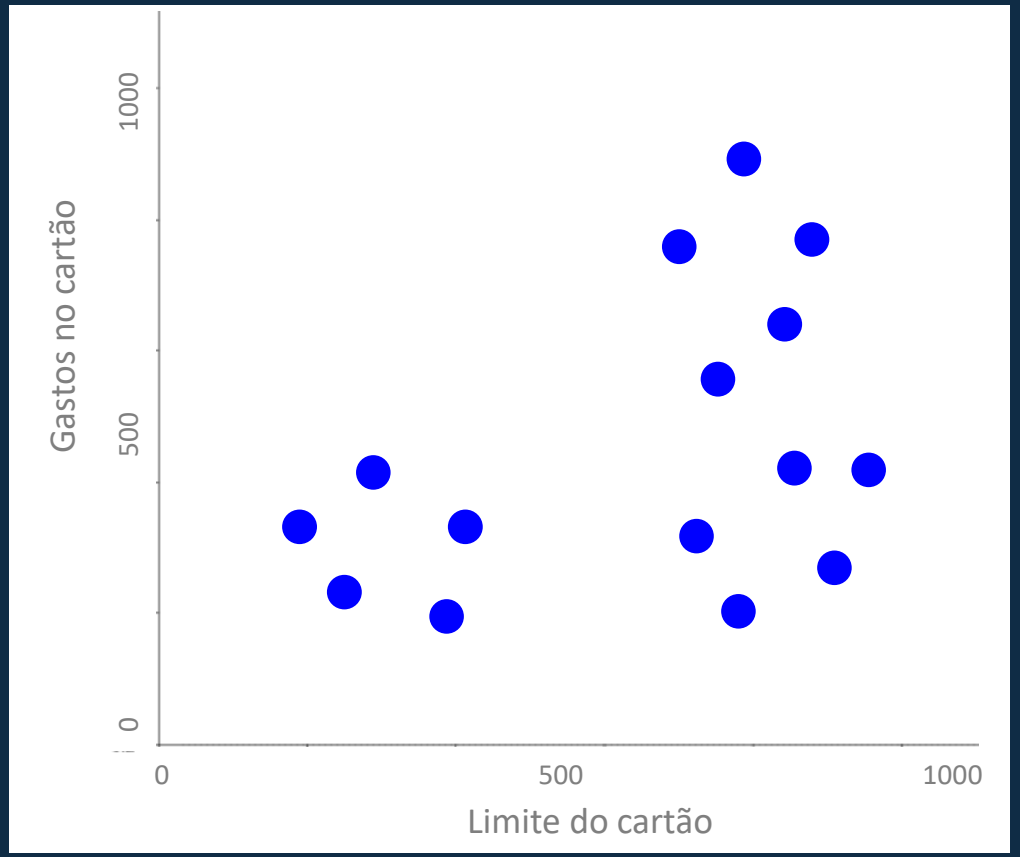
Within Cluster Sum of Squares

$$WCSS = \sum_{i=1}^{k} \sum_{x_i \in k}^{n} d(x_{i,}, c)$$

- Calcula a distância entre cada objeto para seu centroide.

- Variamos o número de centroides e calculamos esta métrica várias vezes para comparar os resultados.

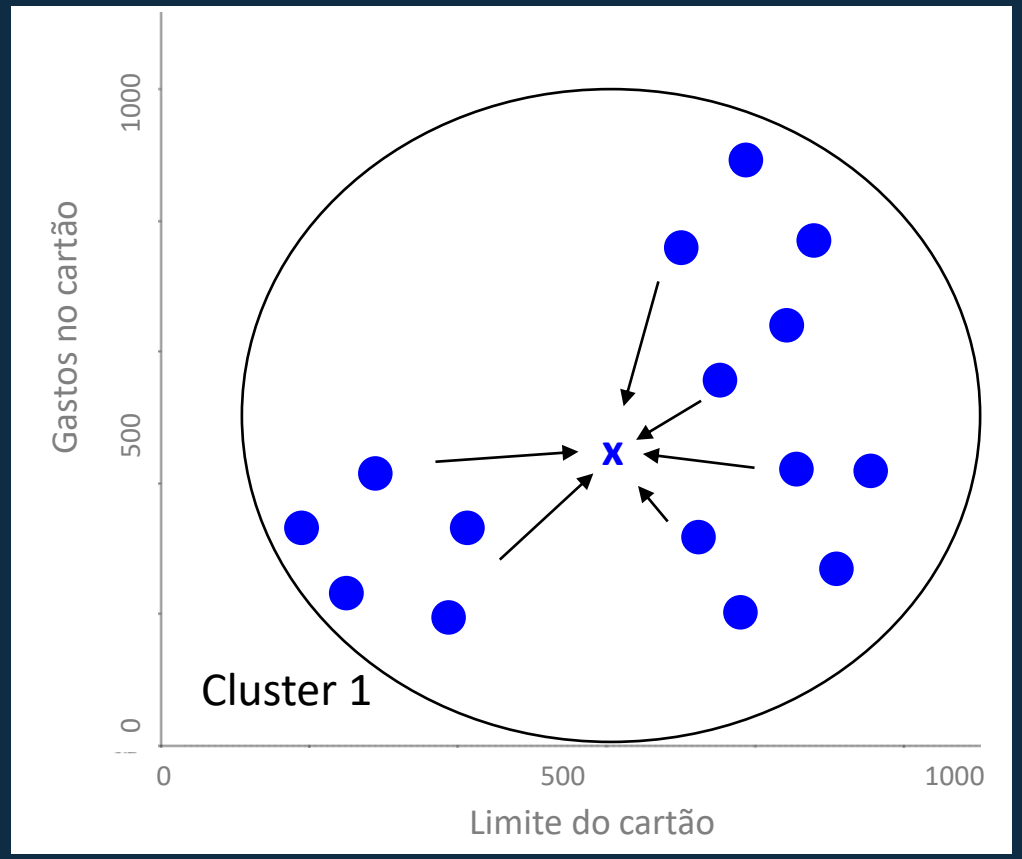
- O resultado desta comparação gera a Curva do Cotovelo (ou, como conhecida no inglês, Elbow Method)



Como definir o número correto de clusters?

Within Cluster Sum of Squares

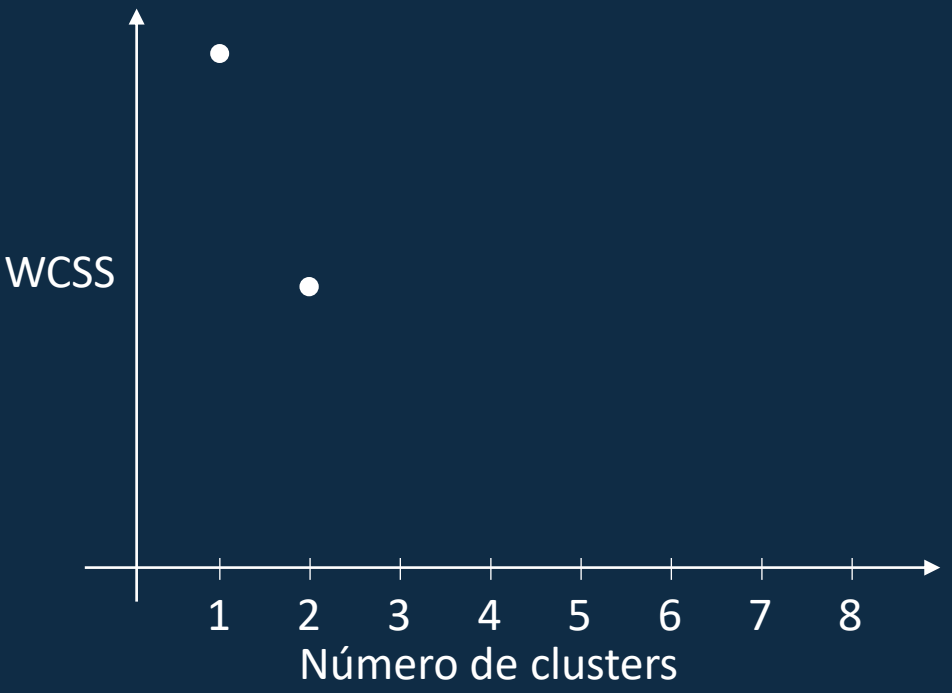
$$WCSS = \sum_{i=1}^{k} \sum_{x_i \in k}^{n} d(x_{i,}, c)$$

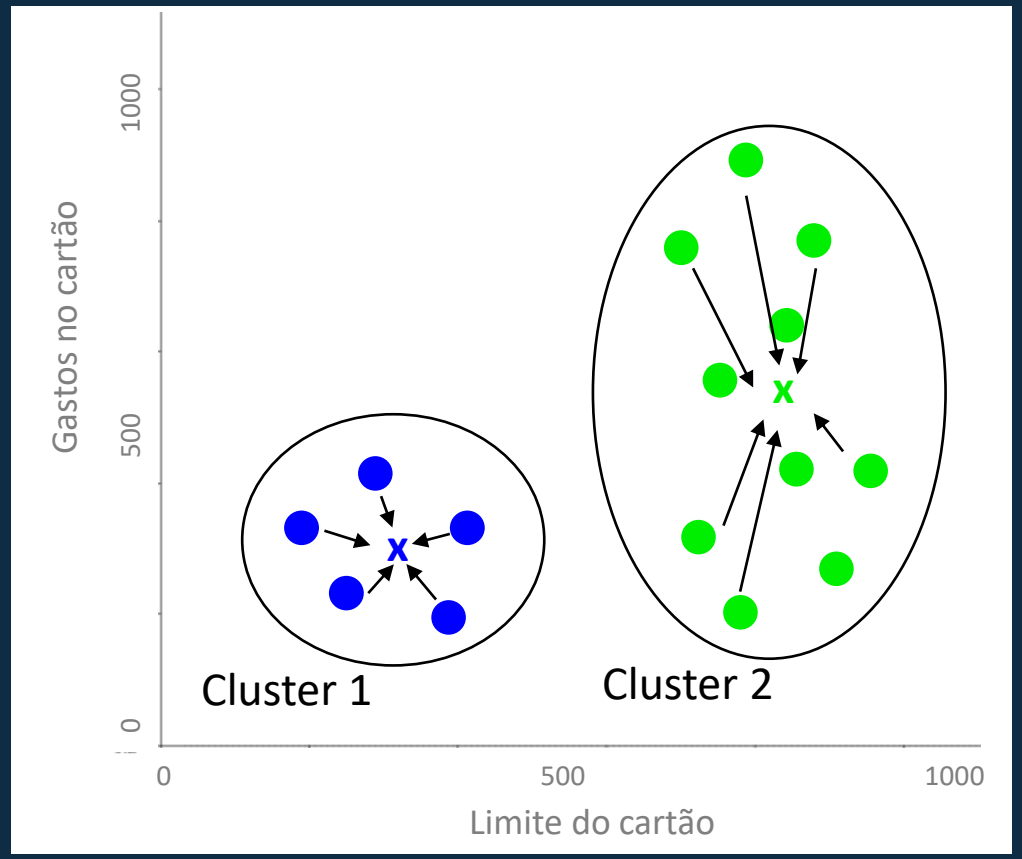


Como definir o número correto de clusters?

Within Cluster Sum of Squares

$$WCSS = \sum_{i=1}^{k} \sum_{x_i \in k}^{n} d(x_{i,}, c)$$

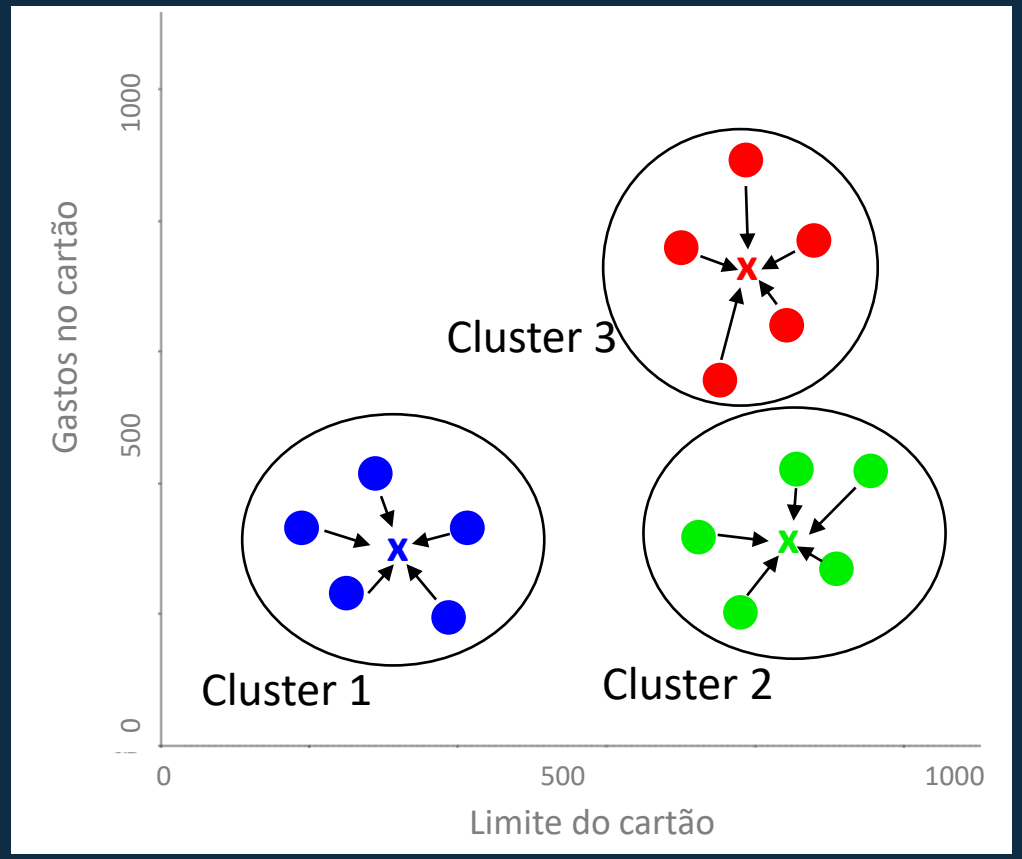




Como definir o número correto de clusters?

Within Cluster Sum of Squares

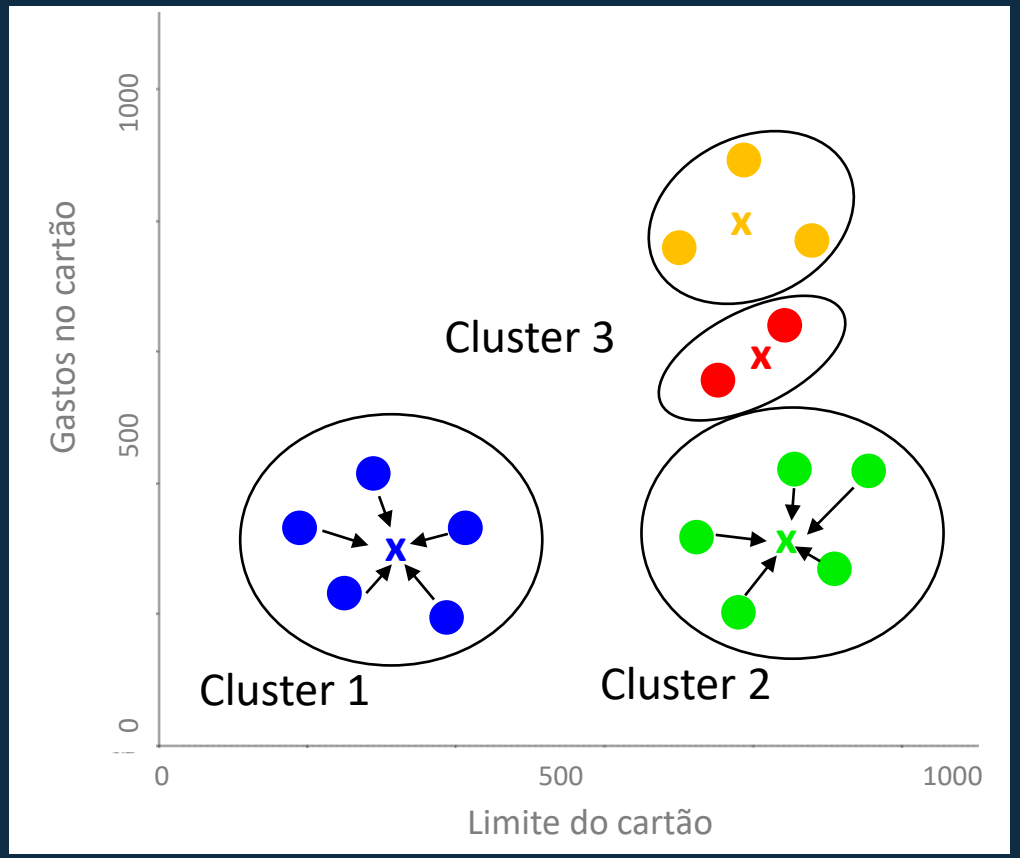
$$WCSS = \sum_{i=1}^{k} \sum_{x_i \in k}^{n} d(x_{i,}, c)$$



Como definir o número correto de clusters?

Within Cluster Sum of Squares

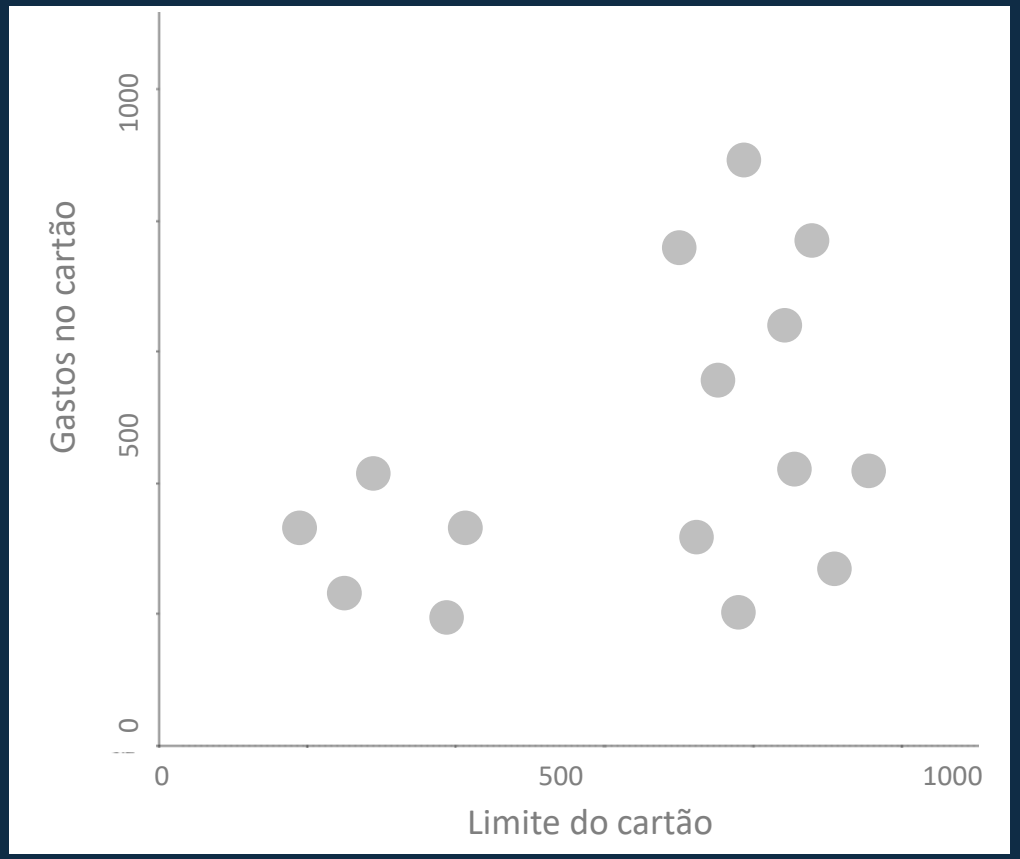
$$WCSS = \sum_{i=1}^{k} \sum_{x_i \in k}^{n} d(x_{i,}, c)$$



Como definir o número correto de clusters?

Within Cluster Sum of Squares

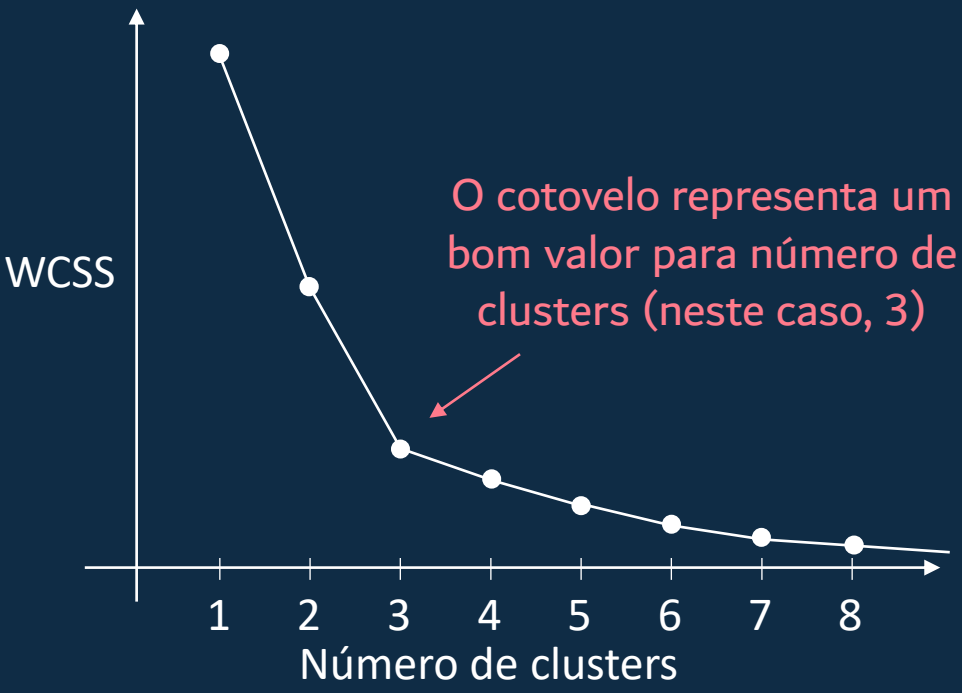
$$WCSS = \sum_{i=1}^{k} \sum_{x_i \in k}^{n} d(x_{i,}, c)$$

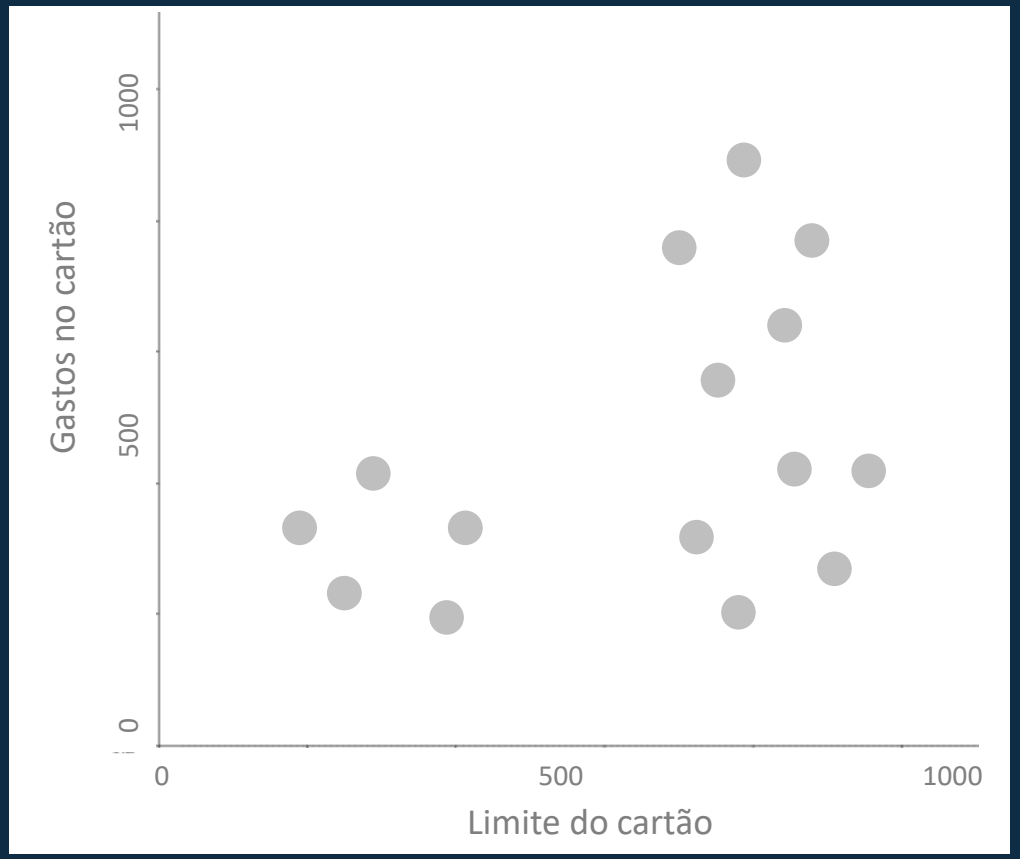


Como definir o número correto de clusters?

Within Cluster Sum of Squares

$$WCSS = \sum_{i=1}^{k} \sum_{x_i \in k}^{n} d(x_{i,}, c)$$





Vantagens

- Algoritmo muito simples de entender;
- Fácil de configurar;
- Rápido e eficiente;
- Funciona muito bem para clusters bem definidos.

Desvantagens

- Não lida bem com outliers (pode gerar clusters incoerentes influenciado por outliers);
- Pode gerar clusters incoerentes dependendo do centroide inicial gerado aleatoriamente;
- Obriga a definir o numero de cluster na inicialização;
- Pode gerar clusters incoerentes em clusters com formatos não globulares

Contornando as desvantagens:

- K-medians, K-medoids
- K-means++
- Elbow method
- Outras abordagens de agrupamento

Para se aprofundar mais

- K-medians, K-medoids
- Funcionamento do K-means++
- Métricas de avaliação intra-cluster, extra-cluster, silhueta etc.

Modelo	Objetivo Resumido
K-means	A partir de um conjunto de dados, o algoritmo cria K grupos pela semelhança dos dados. O número inicial de grupos (K) é definido pelo usuário. O algoritmo utiliza métricas de distância para avaliar a distância entre os pontos. Possui um elemento representativo para cada cluster, chamado de centroide. O objetivo final dele é reduzir a distancia intra-cluster para o menor número possível, com um número de clusters adequado.

Tipo	Agrupamento
Categoria	Não-supervisionado

Premissas para o funcionamento	- Dados precisam estar normalizados (algoritmo baseado em distância); - Dados precisam ser numéricos;	Hiperparâmetros	- N_clusters - Init - N_init - max_iter
Funcionamento detalhado (Steps, informações detalhadas etc.)	O algoritmo inicializa aleatoriamente um número K de centroides e, utilizando cálculos de distância (ex: euclidiana), atribui cada ponto a um dos clusters (o que tiver a menor distância). Posteriormente, o algoritmos inicia um loop que gera um novo centroide a partir da **média entre todos os pontos dentro deste cluster.** Gerado o novo centroide, novamente é avaliada a distância de cada ponto para os centroides e definido o novo cluster. O processo roda novamente até que nenhuma variável mude de cluster (convergência).		**STEPS:** 1. Define-se o número K de Clusters; 2. Aleatoriamente, define-se a posição de cada centroide (centro do cluster); 3. Calcula-se a distância (ex: euclidiana) entre cada ponto e os centroides, definindo qual é o centroide mais próximo; 4. Gera-se um novo centroide a partir da média de todos os pontos definidos em cada cluster; 5. Repete os passos 3 e 4 até que não haja mais movimentação de pontos entre clusters.
Vantagens	- Algoritmo muito simples de entender - Fácil de configurar - Rápido e eficiente	Desvantagens	- Não lida bem com outliers (pode gerar clusters incoerentes influenciado por outliers); - Pode gerar clusters incoerentes dependendo do centroide inicial gerado aleatoriamente; - Obriga a definir o numero de cluster na inicialização;
Como avaliar o desempenho	- Pode-se extrair a métrica ***Sum of Squared Error*** (SSE) para várias opções de número de cluster, encontrando assim o número ideal de clusters com base na Elbow Method (conhecida como curva do cotovelo). - Métricas como distância Intra-cluster e distância inter-cluster ajudam a avaliar a qualidade do agrupamento. - Métrica da silhueta auxilia na identificação dos clusters com menor distancia intra e maior distancia inter clusters	Como corrigir ou compensar desvantagens	• K-medians para combater outliers • K-means++ para inicialiazar os centroides • Elbow method para definir o número de K

Obrigado!

youtube.com/@Tech_dados

linkedin.com/in/itallo-dias/