FIAP

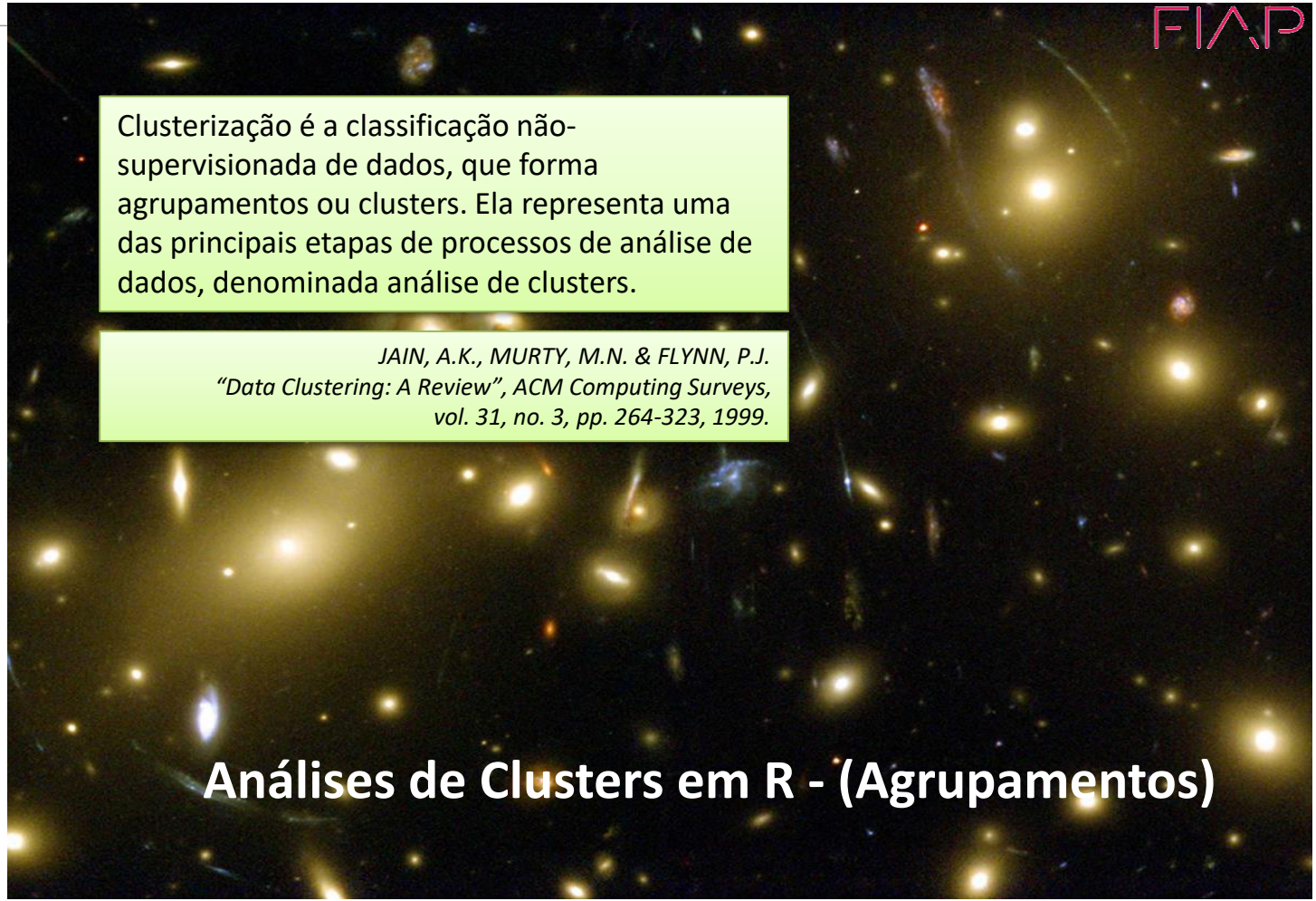
MBA

MBA EM DATA SCIENCE & AI

STATISTICS WITH R

AULA 6

Análise de Cluster



Clusterização é a classificação não-supervisionada de dados, que forma agrupamentos ou clusters. Ela representa uma das principais etapas de processos de análise de dados, denominada análise de clusters.

JAIN, A.K., MURTY, M.N. & FLYNN, P.J.
"Data Clustering: A Review", ACM Computing Surveys,
vol. 31, no. 3, pp. 264-323, 1999.

Análises de Clusters em R - (Agrupamentos)

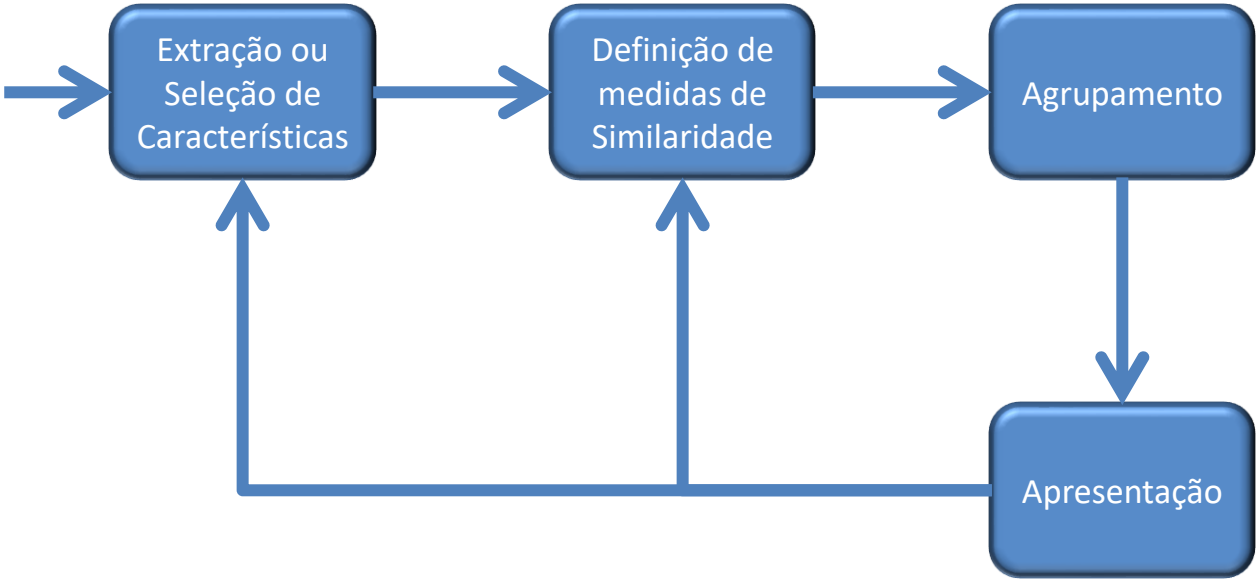
Análise de clustes

- a análise de clusters envolve, portanto, a organização de um **conjunto de padrões** (usualmente representados na forma de vetores de atributos ou pontos em um espaço multidimensional – espaço de atributos) em clusters, de acordo com alguma

  medida de similaridade.

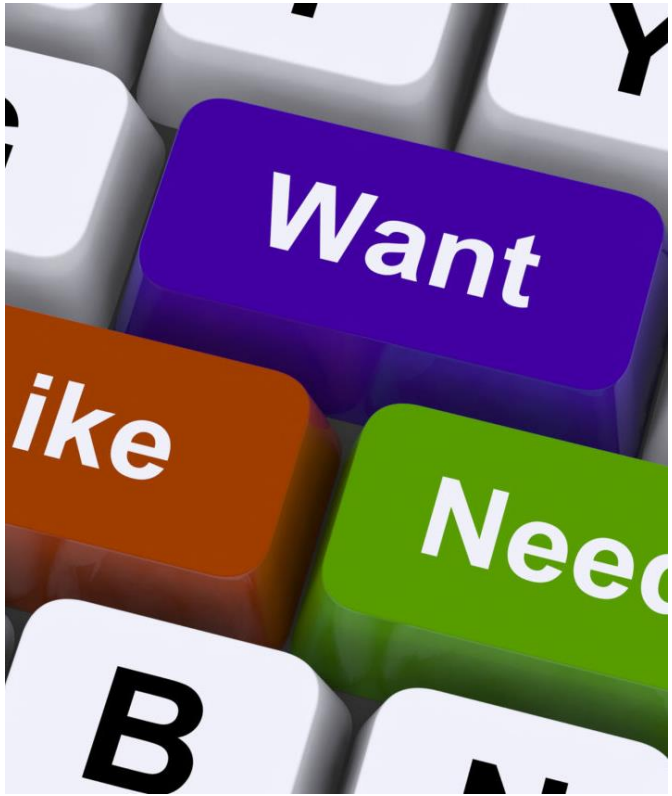
- intuitivamente, padrões pertencentes a um dado **cluster** devem ser mais “similares” entre si do que em relação a padrões pertencentes a outros clusters.

Etapas de um processo de clusterização

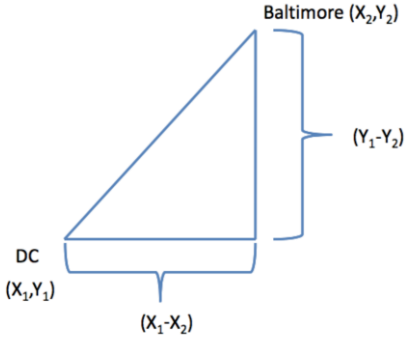


Pré-requisitos

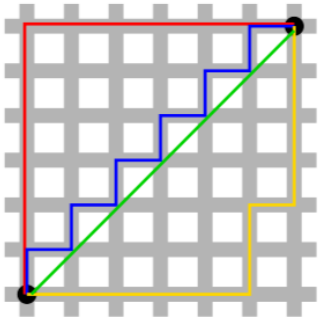
- Como definimos “próximo”?
- Como as observações são agrupadas?
- Como visualizar os agrupamentos?
- Como interpretar os agrupamentos?



K-Médias – Conceito de ‘próximo’



- Definir qual a métrica da distância
  - Distânica Direta
    - Distância euclidiana
    - Distância de Manhattan
  - Binária
    - Distância de Manhatan
  - Contínua – (Correlação)



- Quando houver diferentes métricas
  - Normalizar as comparações de distância / similaridade

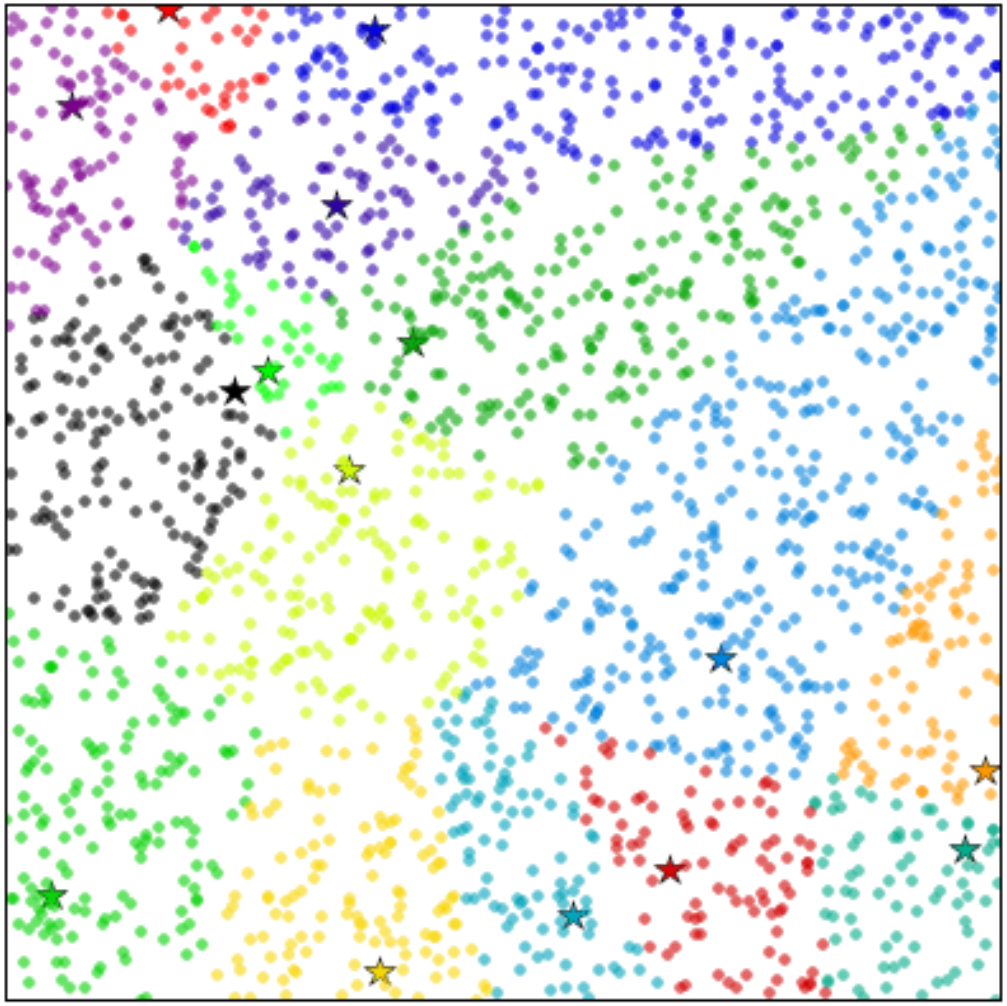
Algoritmos de Clusterização

- Não há uma técnica de clusterização universal
  - Que seja capaz de revelar toda a variedade de estruturas que podem estar presentes em conjuntos de dados multidimensionais;

- Técnicas mais populares:
  - K-Médias
  - Hierárquico

K-Médias / K-Means

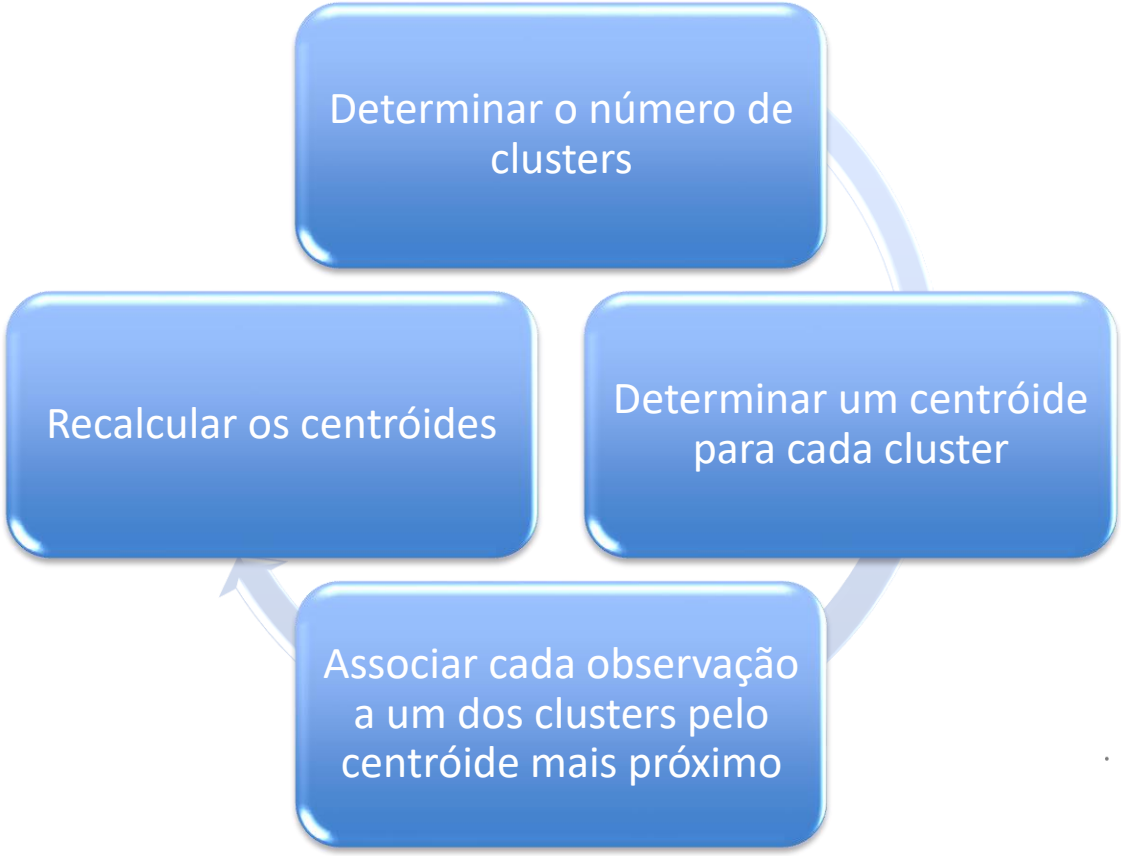
N = 2000, K = 15 Iteration 1



K-Médias / K-Means

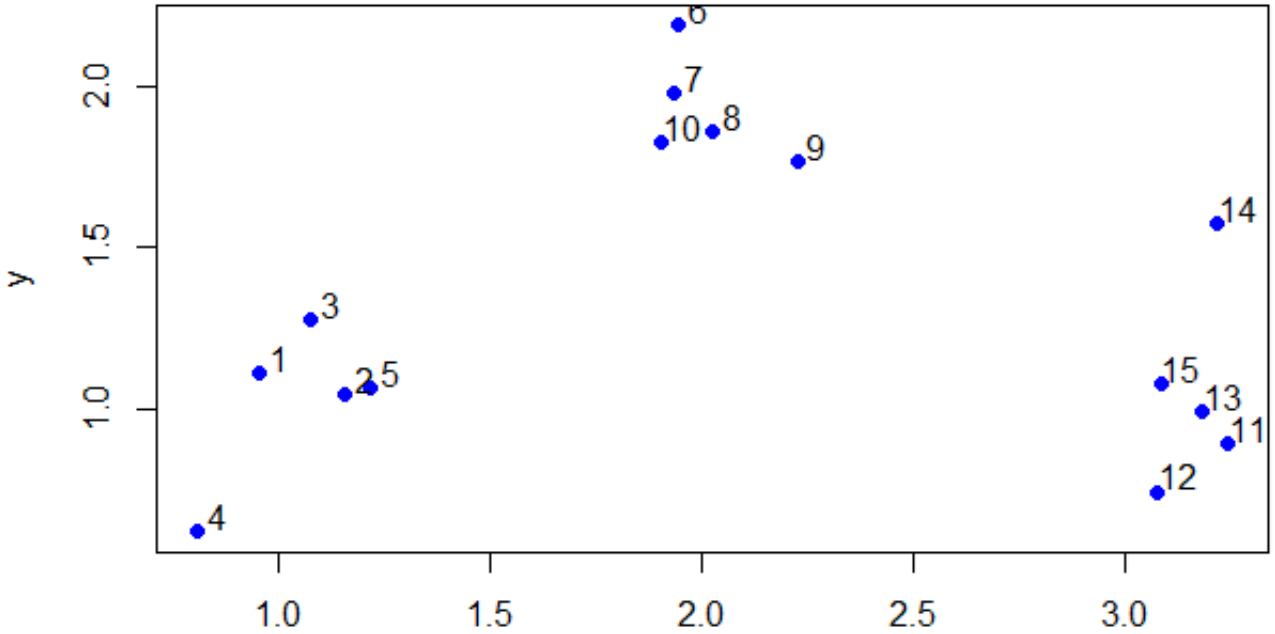
- Trata-se de uma técnica muito muito antiga mas ainda muito útil nos dias de hoje (1956).
- Sumariza grandes volumes de dados, mesmo com alto número de dimensões.
- Possibilita rápida análise de padrões.
  - Os registros (as observações) semelhantes são exibidas próximas umas das outras e um centro geométrico classifica os agrupamentos

Passos principais



Relembrando distribuições

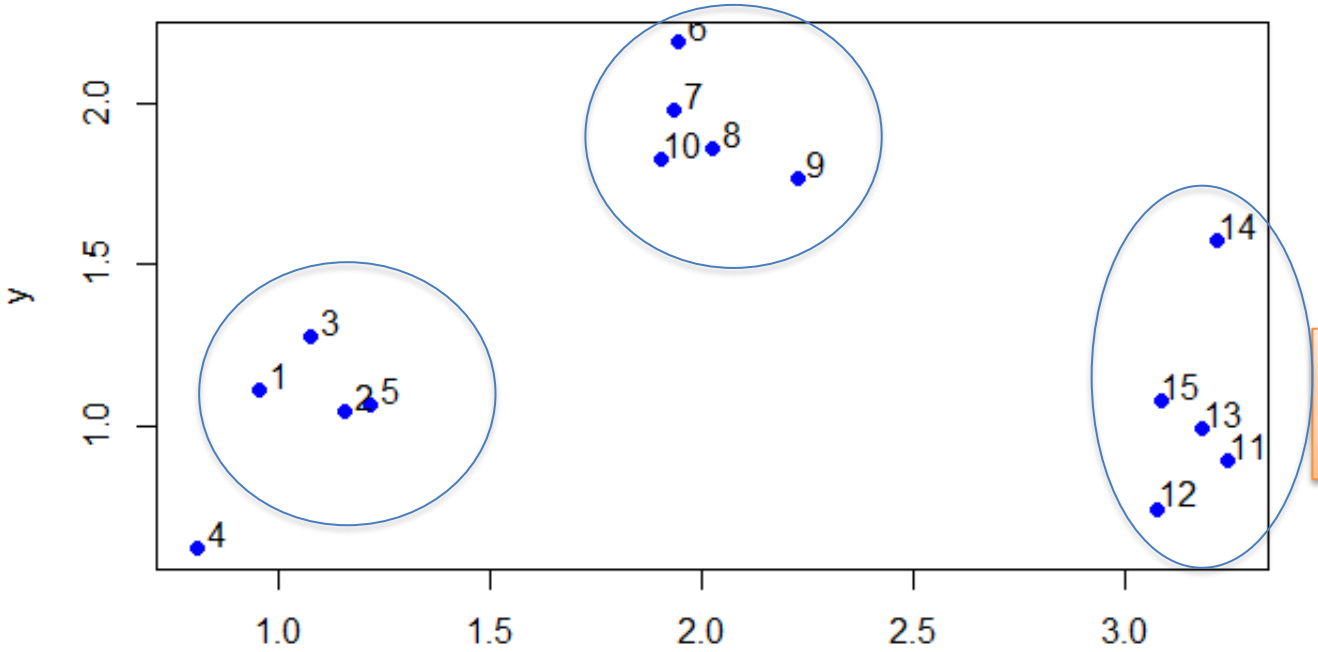
- Vamos fazer um exemplo simples?



Discussão:
O que é preciso para gerar dados como estes?

Relembrando distribuições

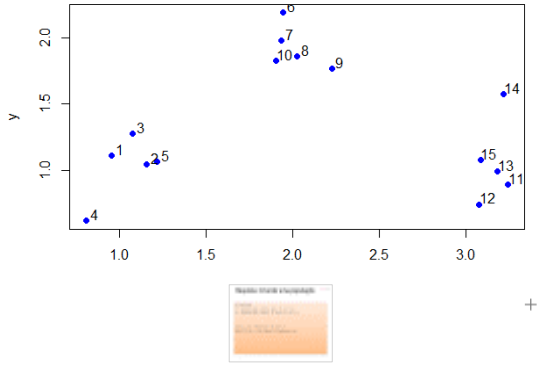
- Vamos fazer um exemplo simples?



Discussão:
O que é preciso para gerar dados como estes?

Distribuição normal e K-Means

- Criar um código que gera uma população amostral com apenas duas dimensões : x e y
- Usar a distribuição normal para gerar uma variação nos dados gerados
  - Sugestão:
    - População com 15 observações
    - 3 centróides
    - 5 observações por centróide
- Essa população deve ser impressa (plot) no plano cartesiano



Como rodar o K-Means

```
modelo = kmeans(x = mydata, centers = centros)
```

- Pergunta:
  - O que é preciso fazer com os dados que temos para passar para o parâmetro x?

Saídas notáveis deste modelo:

cluster – O cluster a que cada observação pertence
centers – as coordenadas de cada centróide (final)
size – Elementos por cluster
iter – Número de iterações para achar a resposta

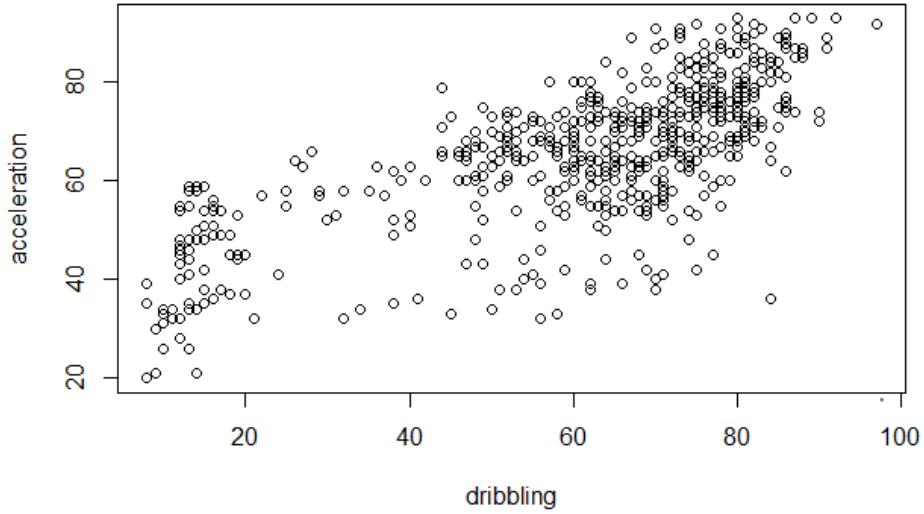
K-Means na prática

Vamos explorar um dataset real

Fifa 18 – Analisado habilidades

- O Dataset possui 17994 observações
  - Filtrar em um dataset menor usando o dplyr que contemple:
    - Atributos name, dribbling, acceleration
    - Apenas os 602 jogadores da liga "Spanish Primera División”

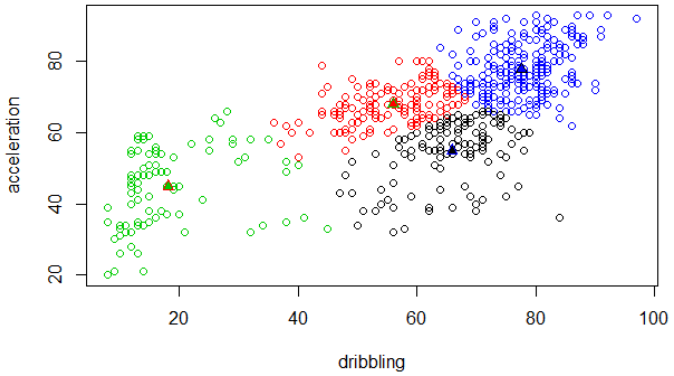
- Plotar o dataset menor:

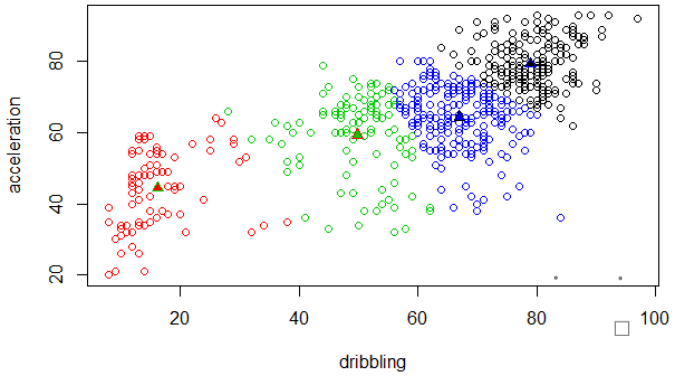


Fifa 2018 – Analisado habilidades

- Rode o k-means com 4 centróides

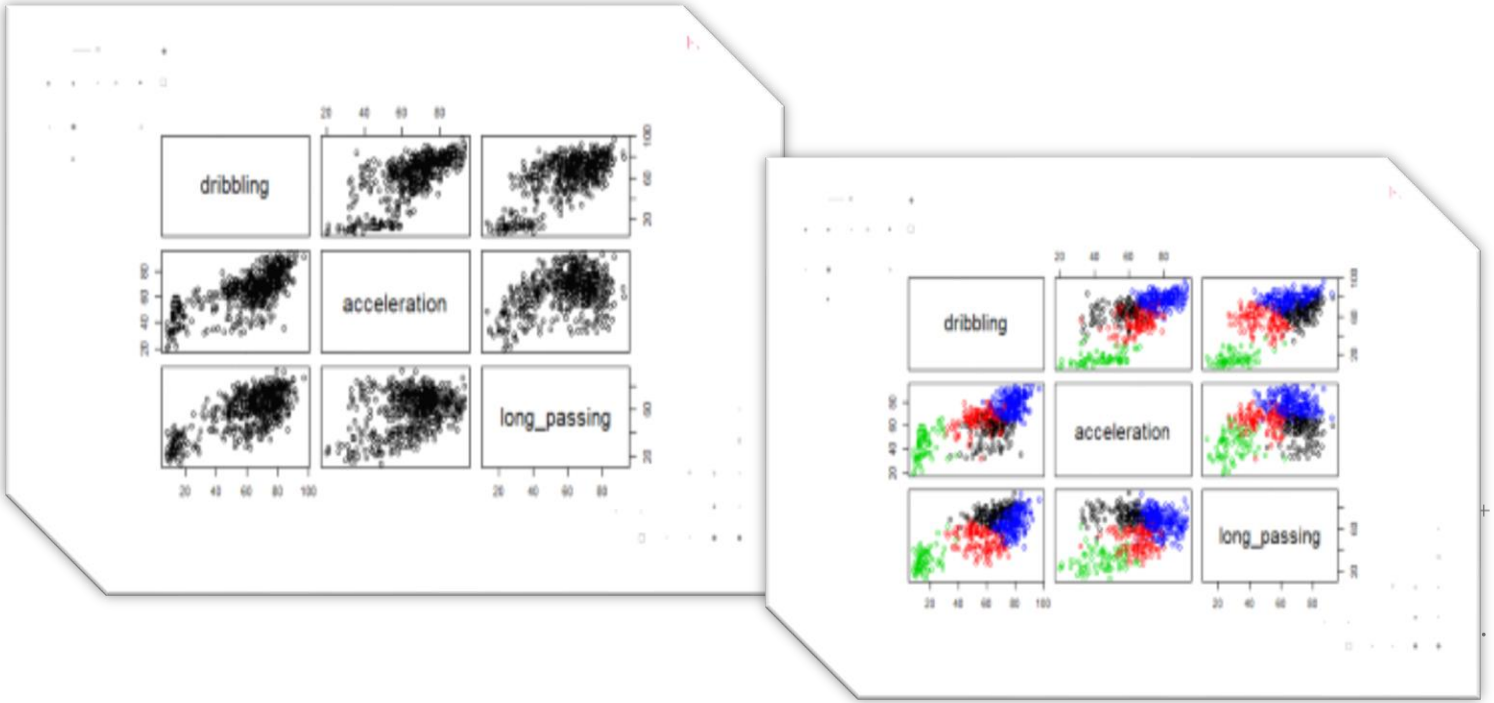
- Observações para discussão:
  - Trata-se de um algoritmo não determinístico
  - O número de clusters depende de análise prévia dos dados, da intuição ou por testes





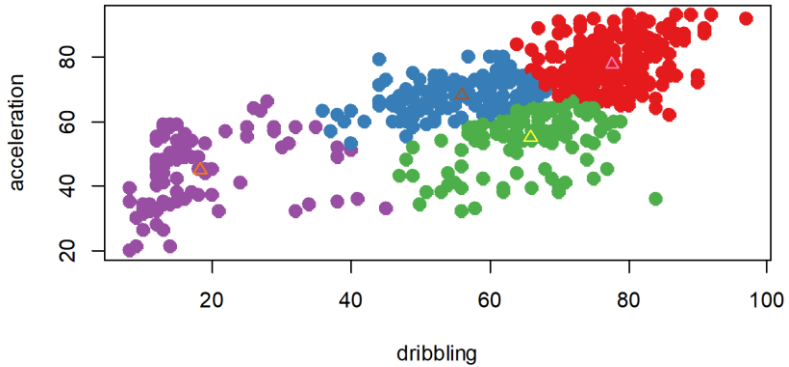
Análise: clusters com mais dimensões

- Análises com 3 dimensões



Avaliação de um modelo K-Means

Avaliação de um modelo K-Means



Medidas total de dispersão:		
totss	(Distância de todos os pontos para todos os pontos. Não muda nunca pela clusterização	=
withinss	Distância de todos os pontos para todos os pontos dentro de um cluster	↓
tot.withinss	Soma de withinss	↓
betweenss	Soma das distâncias de todos os pontos para os demais clusters $k = 1 \rightarrow \nexists; k = pop \rightarrow totss$	↓
$\frac{\text{betweenss}}{\text{totss}}$	$k = 1 \sim 0$ $k = pop \rightarrow 1$	↓

Avaliação de um modelo K-Means

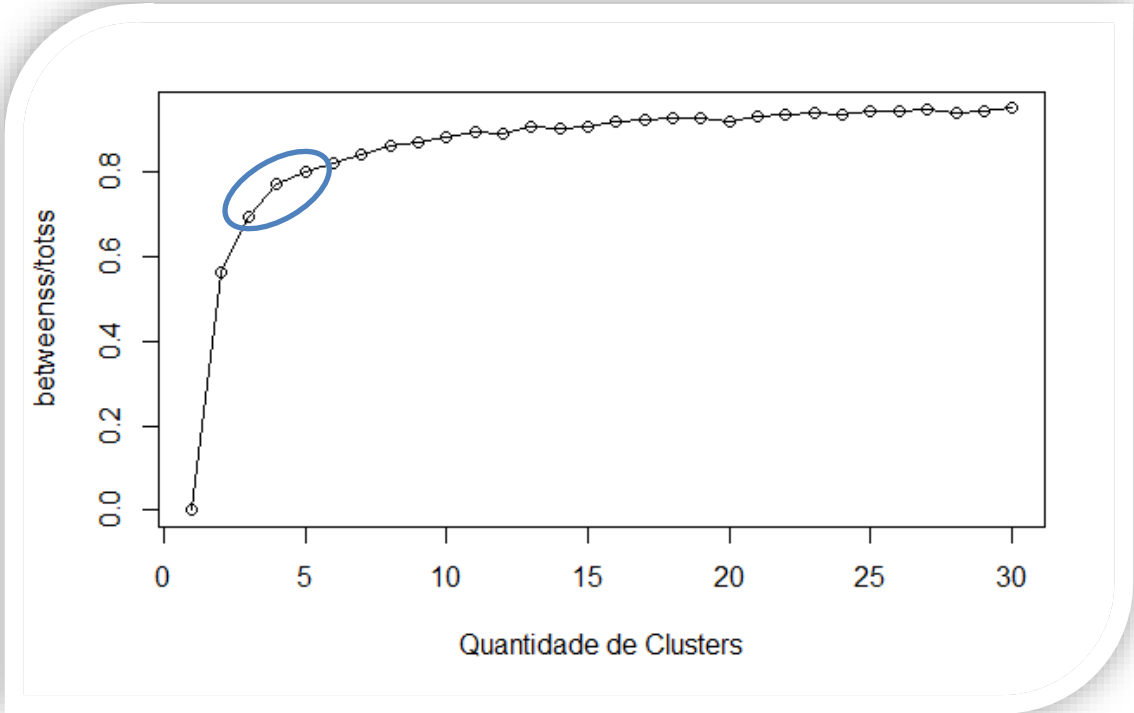
- Experimentos

```
modelo = kmeans(spain.2d[,c('dribbling','acceleration')], centers =
4)
rbind( cbind('betweenss:',  modelo$betweenss),
       cbind('totss:',  modelo$totss),
       cbind('quality:',  modelo$betweenss / modelo$totss ))

plot(spain.2d[,c('dribbling','acceleration')],
     col = modelo$cluster,
     pch = 20, cex = 2)
points(modelo$centers, col = rev(seq_along(modelo$centers)),
       bg=seq_along(modelo$centers), pch = 24, cex =1, lwd = 1)
```

- Avaliar
  - betweenss
  - totss
  - betweenss/totss

Avaliação de um modelo K-Means

- Identificação de um número de clusters



Alternativa k-Medóides

Biblioteca cluster

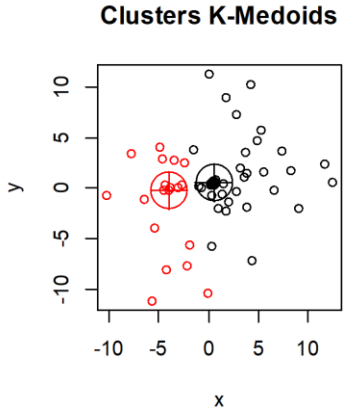
```
library(cluster)
```

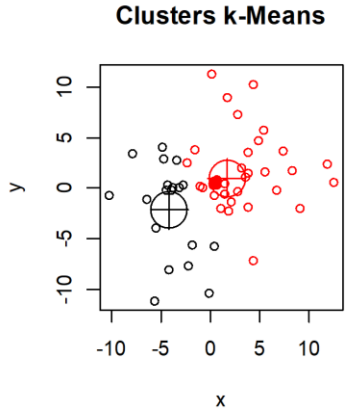
K-Medóides

- Minimiza a soma das distâncias de todos os pontos de um cluster para o centro do cluster
- Determinístico

K-Means

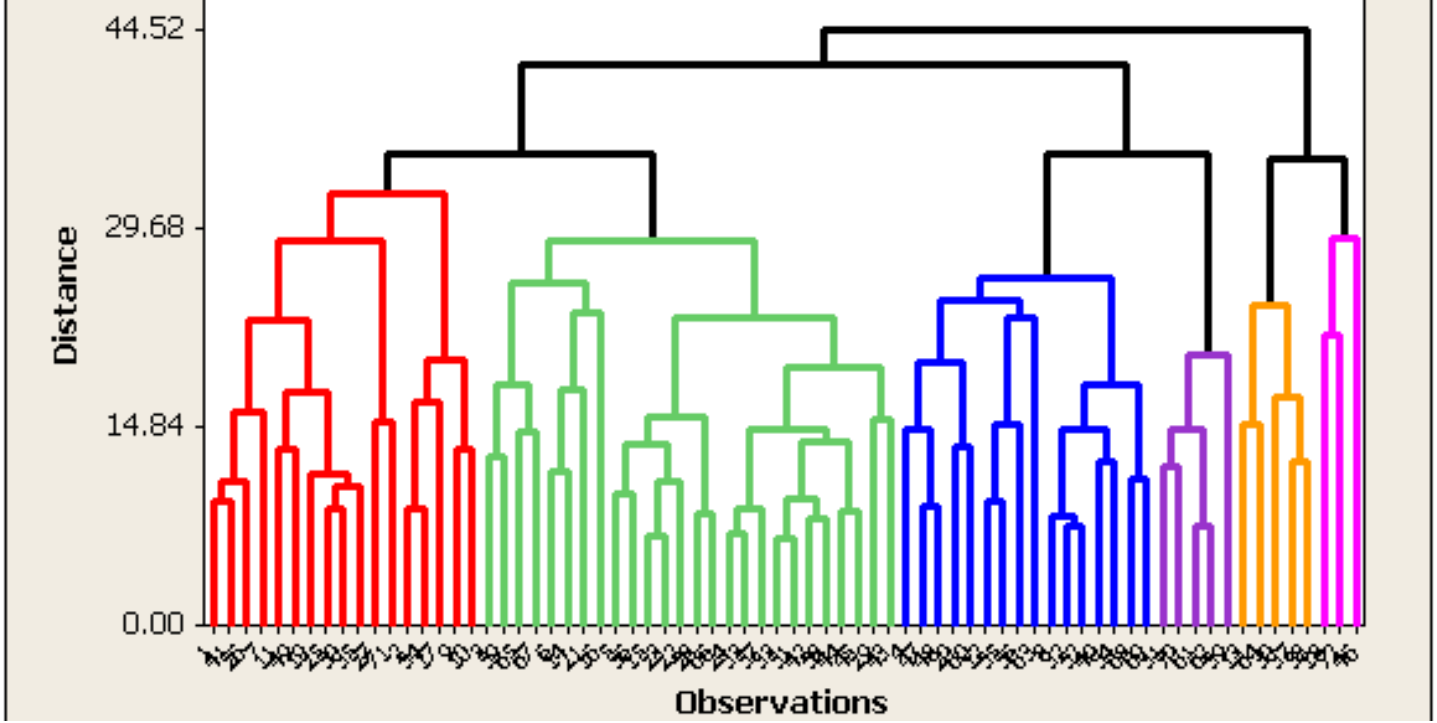
- Minimiza a distância de todos os pontos para os centros de demais clusters
- Não determinístico





Alternativa k-Medóides

```
## K-Medoids

library(cluster)
x <- rbind(matrix(rnorm(100, mean = 0.5, sd = 4.5), ncol =
2),
           matrix(rnorm(100, mean = 0.5, sd = 0.1), ncol =
2))
colnames(x) <- c("x", "y")

modelo1 <- pam(x,2)
modelo2 <- kmeans(x, 2)


par(mfrow=c(1,2))

plot(x, col = modelo1$clustering, main="Clusters k-Medoids")
points(modelo1$medoids, col = 1:3, pch = 10, cex = 4)

plot(x, col = modelo2$cluster, main="Clusters k-Means")
points(modelo2$centers, col = 1:3, pch = 10, cex = 4)
```



Dendrogramas

E análises de clusters hierárquicos

Dendrograma – Características

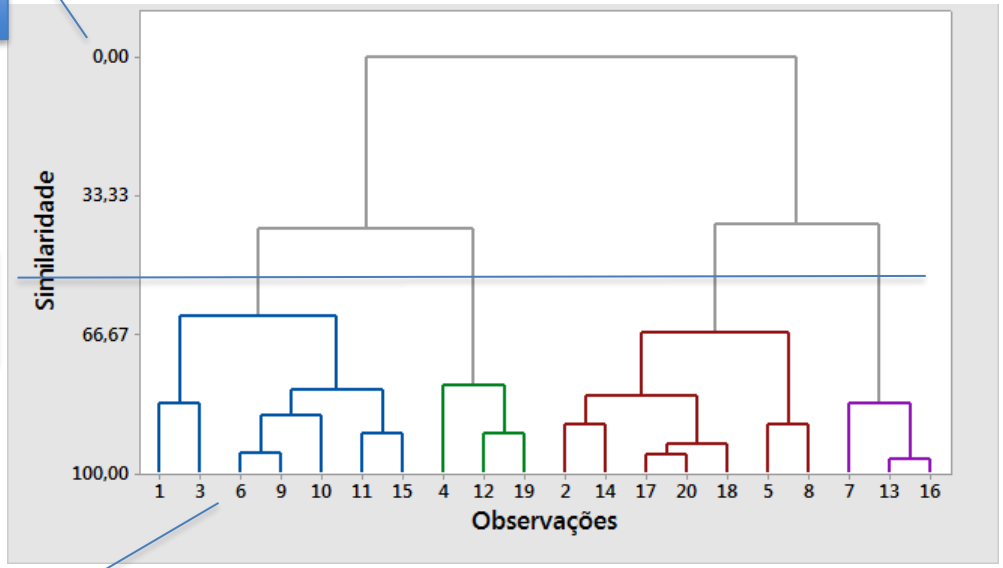
- Método de **clusterização hierárquica** ascendente
- Mostra a evolução dos clusters
- Indicado para populações ‘menores’
- O corte por semelhança/distância indica a quantidade de clusters.

Interpretando o dendrograma

Distância ou similaridade

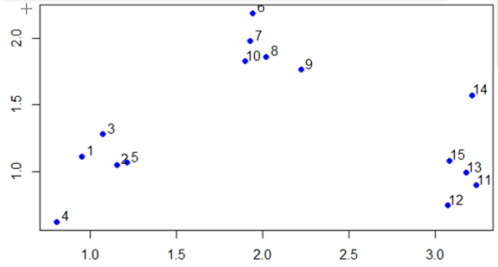
Corte dos clusters

Lista de obervações



Criando um dendrograma simples

Mesmos dados do k-Means

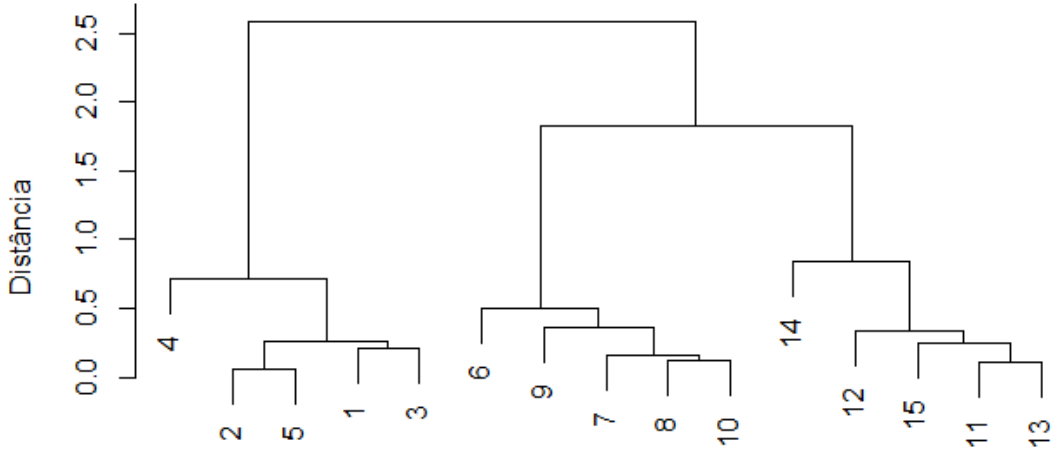



```
set.seed(1909)
x <- rnorm(15, mean = rep(1:3, each = 5), sd = 0.2)
y <- rnorm(15, mean = rep(c(1, 2), each = 5), sd = 0.2)
mydata <- data.frame(x, y)
```

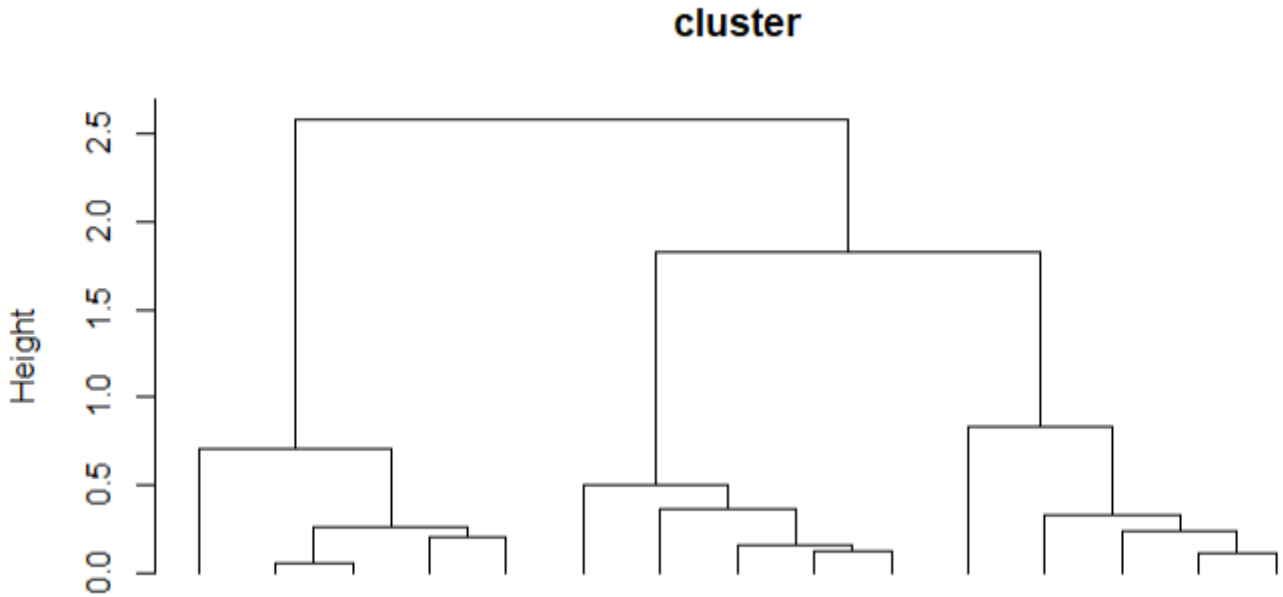
Exemplo de cálculo das distâncias:

```
dist(mydata[1:7,])
dist(mydata[1:7,], method = "manhattan")
```

Dendrograma simples



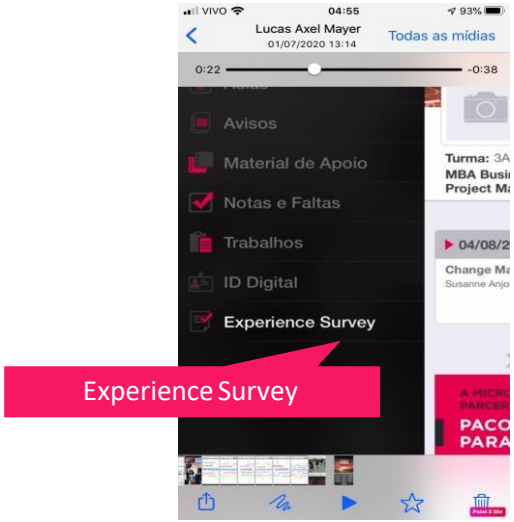
Dendrograma com o Plotly

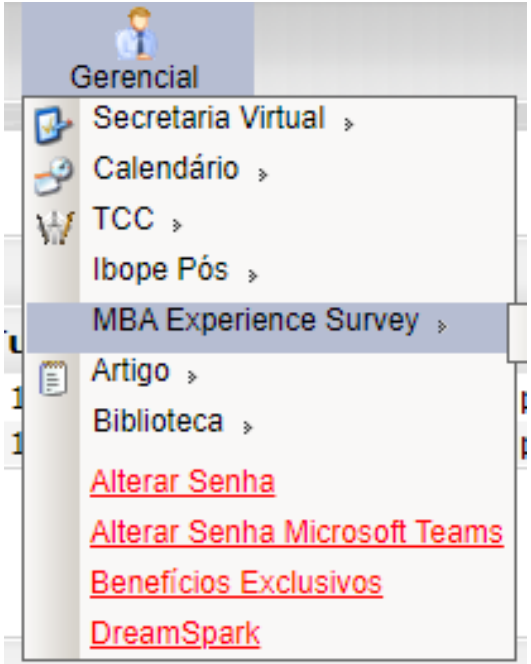


O que você achou da aula de hoje?

Pelo aplicativo da FIAP

(Entrar no FIAPP, e no menu clicar em Experience Survey)





OBRIGADO

in/lafphd

FIAP MBA+


FIAP