

# Relatório

Douglas Cardoso - 11766990

8/26/2021

## Exercício

Obter o dendrograma considerando o método do vizinho mais próximo e matriz de distância Euclidiana para as variáveis padronizadas.

## Tabela

```
tibble::tibble(  
  nome = c('MIT', 'UColombia', 'UChicago', 'Michigan', 'Oxford', 'USP'),  
  num_alunos = c(11067, 25045, 13557, 43147, 19720, 65711),  
  num_funcionarios = c(2982, 6189, 2449, 6809, 6750, 5582),  
  num_estudantes_estrangeiros = c(3717, 8105, 3379, 7527, 7353, 2086),  
  anos = c(157, 264, 162, 201, 922, 191)) -> universidades  
  
cols <- c("Nome",  
  "Nº alunos",  
  "Nº funcionários",  
  "Nº estudantes estrangeiros",  
  "Anos de existência")  
  
knitr::kable(universidades, col.names = cols)
```

Nome	Nº alunos	Nº funcionários	Nº estudantes estrangeiros	Anos de existência
MIT	11067	2982	3717	157
UColombia	25045	6189	8105	264
UChicago	13557	2449	3379	162
Michigan	43147	6809	7527	201
Oxford	19720	6750	7353	922
USP	65711	5582	2086	191

## Padronização das variáveis

```
universidades %>%  
  recipes::recipe(nome ~ .) %>%  
  recipes::step_normalize(recipes::all_predictors()) %>%  
  recipes::prep() %>%  
  recipes::bake(new_data = NULL) %>%  
  dplyr::relocate(nome) %>%  
  knitr::kable(col.names = cols)
```

Nome	Nº alunos	Nº funcionários	Nº estudantes estrangeiros	Anos de existência
MIT	-0.8873232	-1.1129173	-0.6347628	-0.5318651
UColombia	-0.2219557	0.5511401	1.0593108	-0.1743181
UChicago	-0.7687966	-1.3894819	-0.7652544	-0.5151573
Michigan	0.6397184	0.8728475	0.8361625	-0.3848364
Oxford	-0.4754313	0.8422334	0.7689864	2.0244288
USP	1.7137884	0.2361782	-1.2644424	-0.4182520

## Matriz de Distância Euclideana

```
universidades %>%
  recipes::recipe(nome ~ .) %>%
  recipes::step_normalize(recipes::all_predictors()) %>%
  recipes::prep() %>%
  recipes::bake(new_data = NULL) %>%
  dplyr::select(-nome) %>%
  dist(method = "euclidean") %>%
  as.matrix() -> euclidean_matrix

rownames(euclidean_matrix) <- colnames(euclidean_matrix) <- universidades$nome

euclidean_matrix[upper.tri(euclidean_matrix, diag = TRUE)] <- ''

euclidean_matrix %>%
  as.data.frame() %>%
  knitr::kable()
```

	MIT	UColombia	UChicago	Michigan	Oxford	USP
MIT						
UColombia	2.49189205411332					
UChicago	0.328395661642587	2.74048495224907				
Michigan	2.90866917412834	0.969582922566073	3.11184317066039			
Oxford	3.53516891629171	2.25116886291144	3.72424599063871	2.6558553199914		
USP	2.99920724338652	3.05051455728086	3.01074373153701	2.44389687392633	3.90659801919811	

## Dendograma

```
universidades %>%
  recipes::recipe(nome ~ .) %>%
  recipes::step_normalize(recipes::all_predictors()) %>%
  recipes::prep() %>%
  recipes::bake(new_data = NULL) %>%
  dplyr::select(-nome) %>%
  dist(method = "euclidean") %>%
  hclust(method = 'single') -> clusters

clusters %>%
  as.dendrogram() %>%
  dendextend::set_labels(universidades$nome[clusters$order]) %>%
```

```
dendextend::set('branches_k_color') -> dendograma
```

(a) Quais as duas universidades mais parecidas?

```
dendograma %>%
  dendextend::cut_lower_fun(h = 1) -> ex
```

```
ex[[1]]
```

```
## [1] "MIT"      "UChicago"
```

(b) Considerando 4 grupos. Quais universidades estão em cada grupo.

```
dendograma %>%
  plot()
```

```
dendograma %>%
  dendextend::rect.dendrogram(k = 4, border = 2, lty = 5, lwd = 2)
```

