Métodos Multivariados de Análise de Dados*

4ª Atividade

Alberson da Silva Miranda

22 de outubro de 2024

*Código disponível em https://github.com/albersonmiranda/analise_multivariada.

Índice

1 Exercício 1 **3**
1.1 Análise gráfica 3D dos dados e uma análise de estatística descritiva 3
1.2 Análise de cluster 5
1.3 Anova 10
1.4 Método de K-means 13

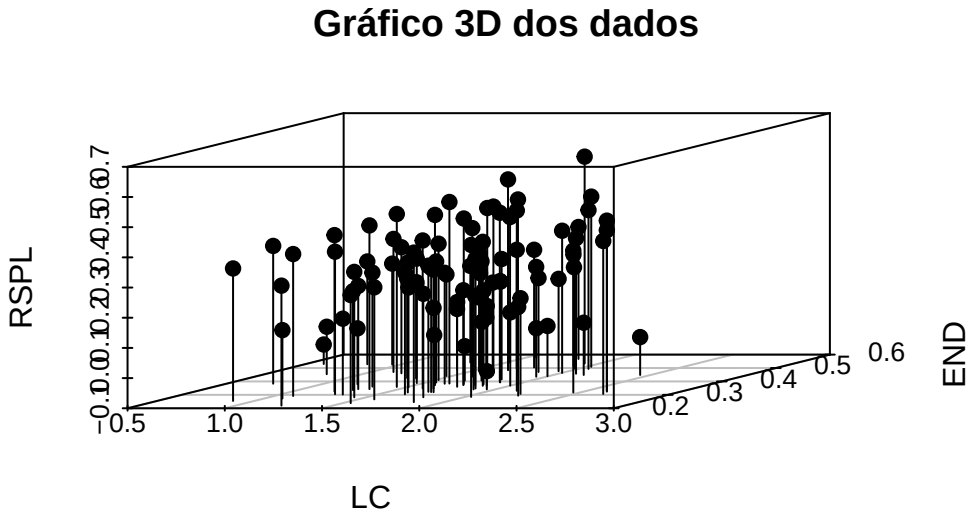
2 Exercício 2 **17**

1 Exercício 1

1.1 Análise gráfica 3D dos dados e uma análise de estatística descritiva

```
# Carregando os dados
dados <- readxl::read_excel("data-raw/clustering/cluster.xlsx")

# plot 3D
scatterplot3d::scatterplot3d(dados[, 2:4], pch = 19, type = "h", main = "Gráfico 3D dos dados")
```

```
Warning: Unknown or uninitialised column: `color`.
```

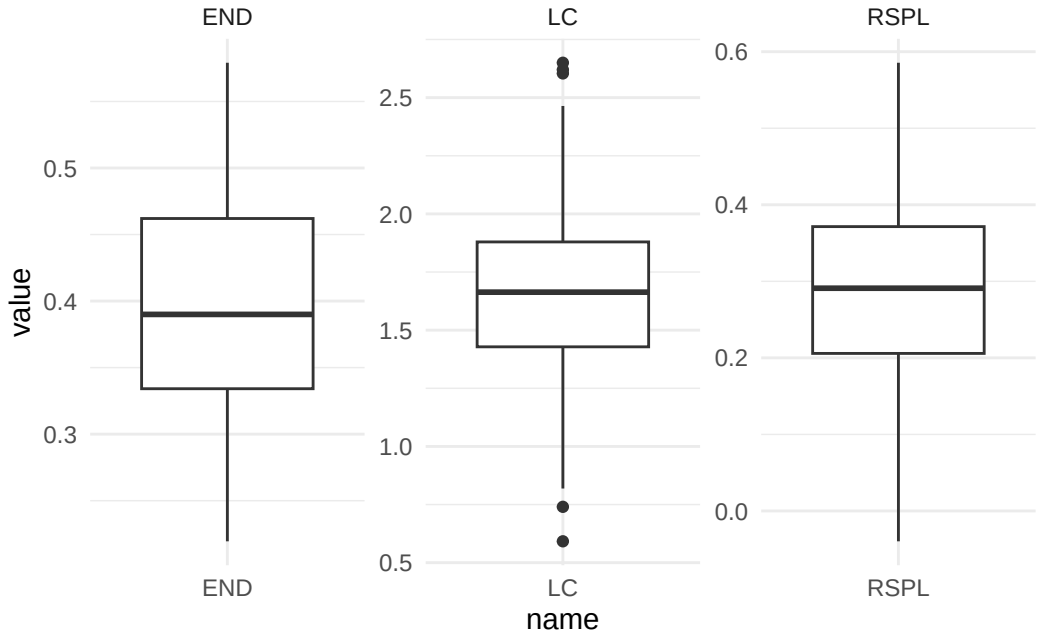



```
# estatísticas descritivas
summary(dados)
```

```
    Empresa             LC              END              RSPL         
 Min.   :  1.00   Min.   :0.5919   Min.   :0.2195   Min.   :-0.03963  
 1st Qu.: 25.75   1st Qu.:1.4283   1st Qu.:0.3341   1st Qu.: 0.20582  
```

```
 Median : 50.50   Median :1.6634   Median :0.3900   Median : 0.29101
 Mean   : 50.50   Mean   :1.6504   Mean   :0.3989   Mean   : 0.27810
 3rd Qu.: 75.25   3rd Qu.:1.8794   3rd Qu.:0.4621   3rd Qu.: 0.37147
 Max.   :100.00   Max.   :2.6500   Max.   :0.5791   Max.   : 0.58556
```

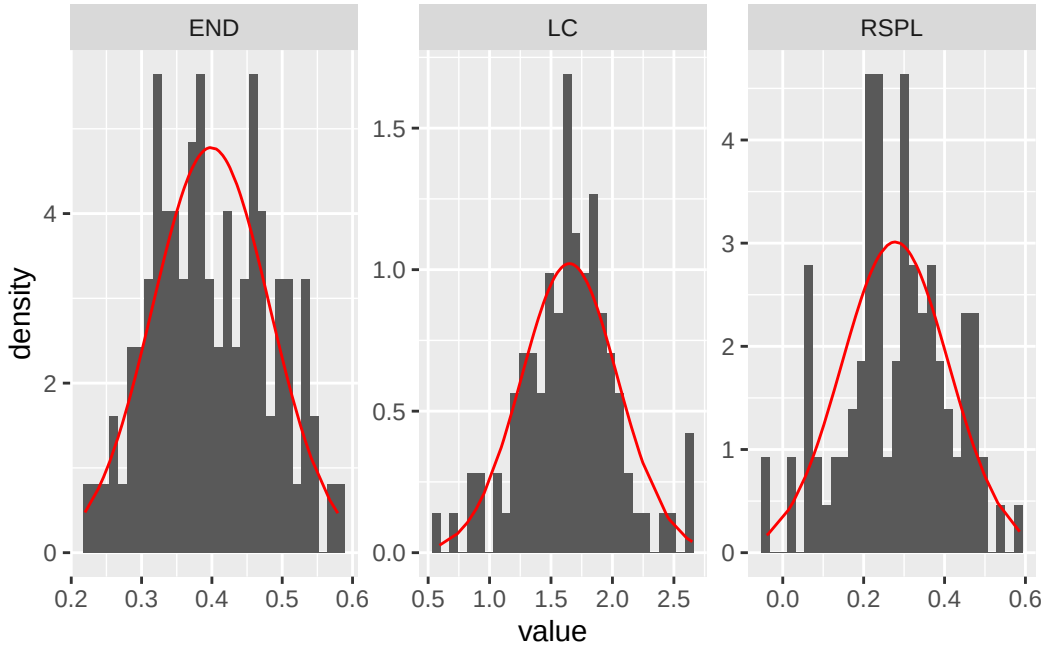
```r
# boxplot
dados_long <- dados |>
  tidyr::pivot_longer(-Empresa)

dados_long |>
  ggplot(aes(x = name, y = value)) +
  geom_boxplot() +
  facet_wrap(~name, scales = "free") +
  theme_minimal()
```




```r
# histograma com linha da normal teórica
dados_long |>
  ggplot(aes(x = value)) +
  geom_histogram(aes(y = after_stat(density))) +
  facet_wrap(~name, scales = "free") +
  geom_line(
    aes(
      y = dnorm(
        value,
```

```
        mean = tapply(value, name, mean, na.rm = TRUE)[PANEL],
        sd = tapply(value, name, sd, na.rm = TRUE)[PANEL]
      )
    ),
    color = "red"
  )
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



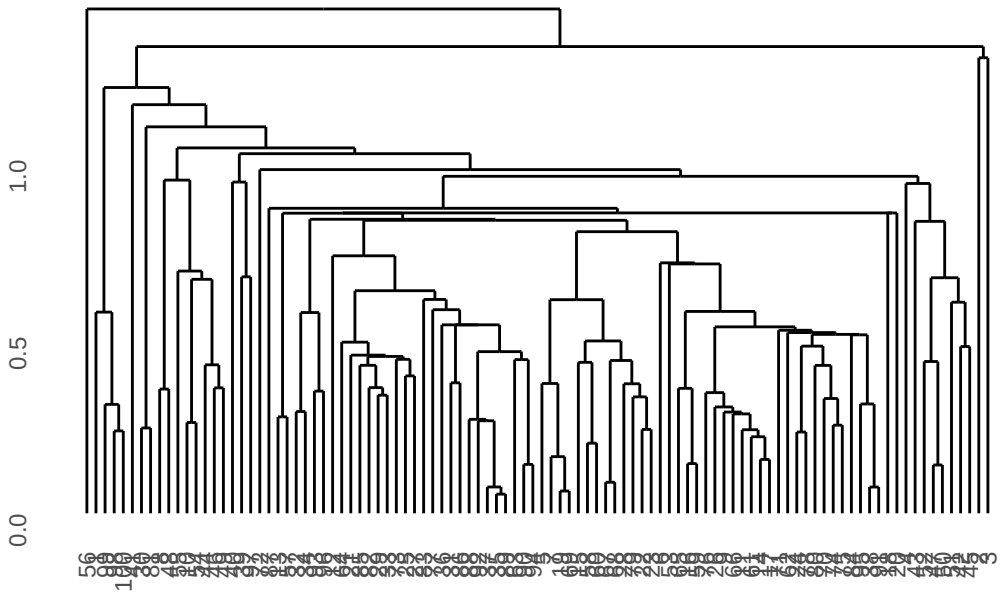
1.2 Análise de cluster

Com *single linkage*, o dendograma não parece muito promissor, com o primeiro *split* já não trazendo muito ganho de informação:

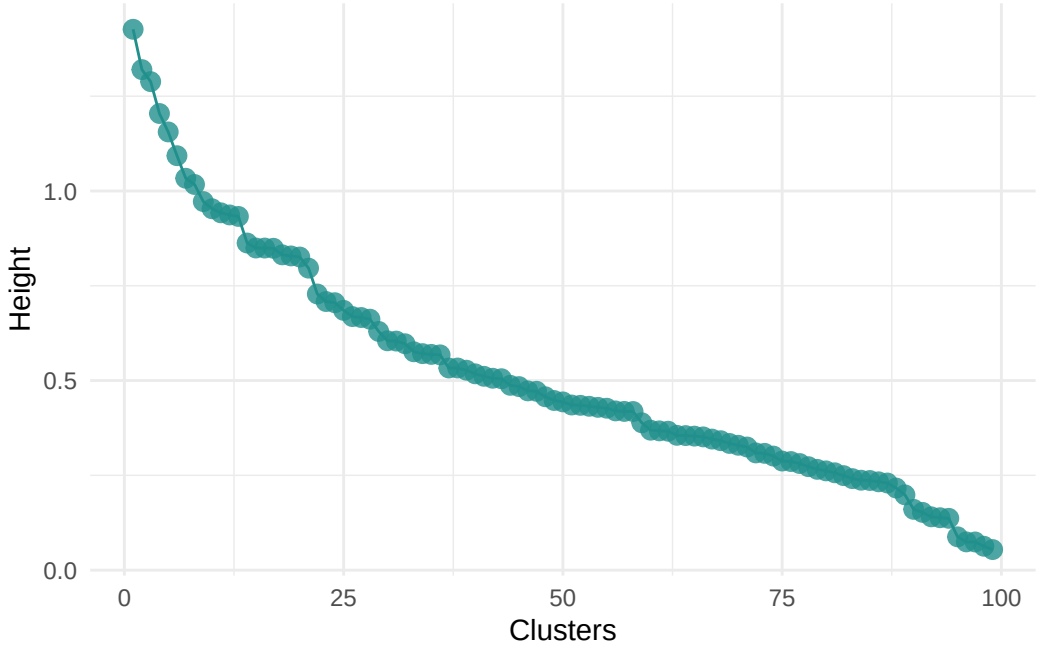
```
# Carregando os pacotes
library(mlr3verse)
```

```
Loading required package: mlr3
```

```r
library(mlr3cluster)

# normalizando os dados
dados_scale <- as.data.frame(scale(dados[, 2:4]))

# Criando tarefa
task <- TaskClust$new(id = "cluster", backend = dados_scale)

# criando learner
learner <- lrn(
  "clust.hclust",
  distmethod = "euclidean",
  method = "single"
)

# treinando o modelo
model <- learner$train(task)

# plotando dendograma
autoplot(model)
```

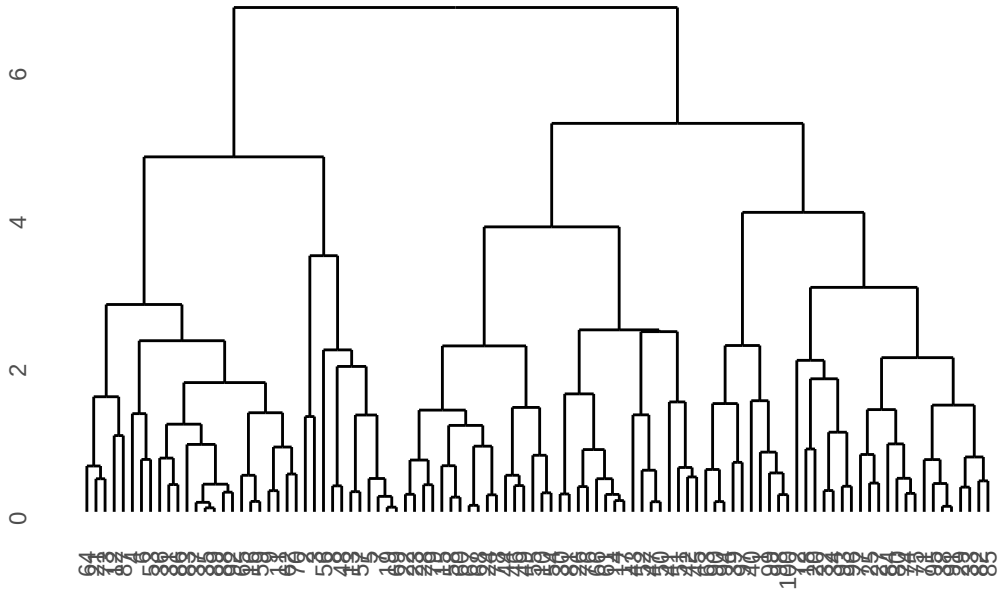



```r
# plotando altura x número de clusters
autoplot(model, type = "scree")
```

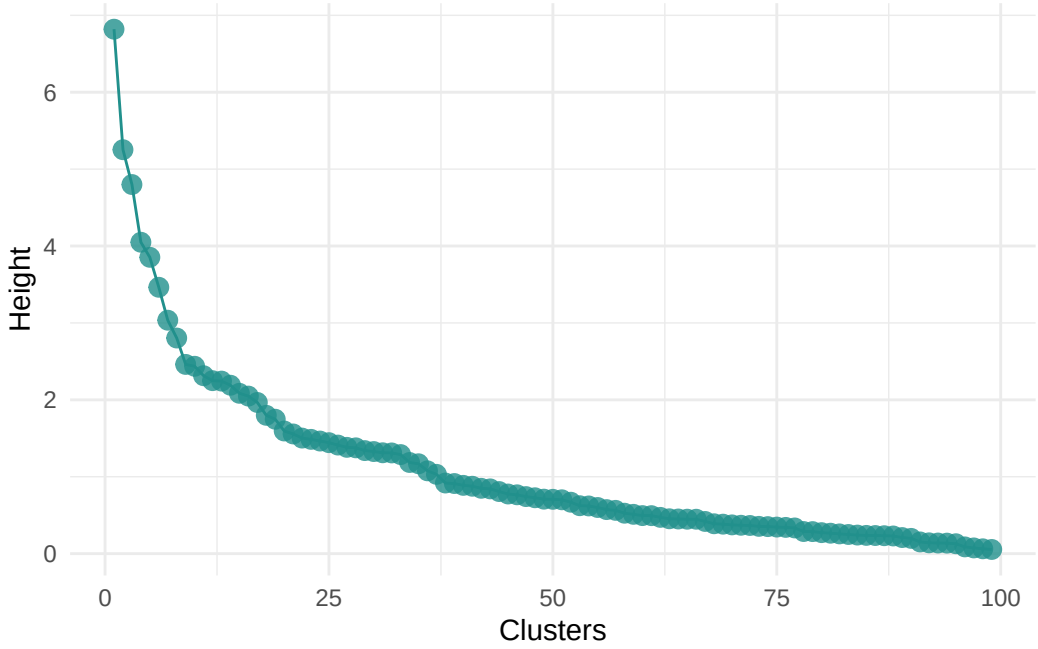


Já com o método *complete*, o dendograma é mais coerente, sugerindo 3 clusters:

```
# criando learner
learner <- lrn(
  "clust.hclust",
  distmethod = "euclidean",
  method = "complete"
)

# treinando o modelo
model <- learner$train(task)

# plotando dendograma
autoplot(model)
```

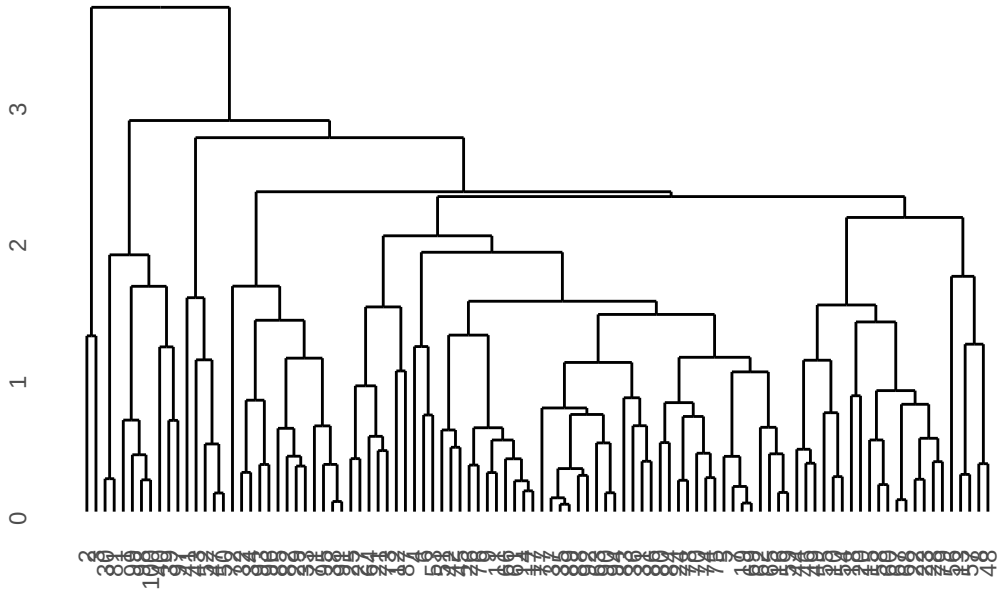



```
# plotando altura x número de clusters
autoplot(model, type = "scree")
```

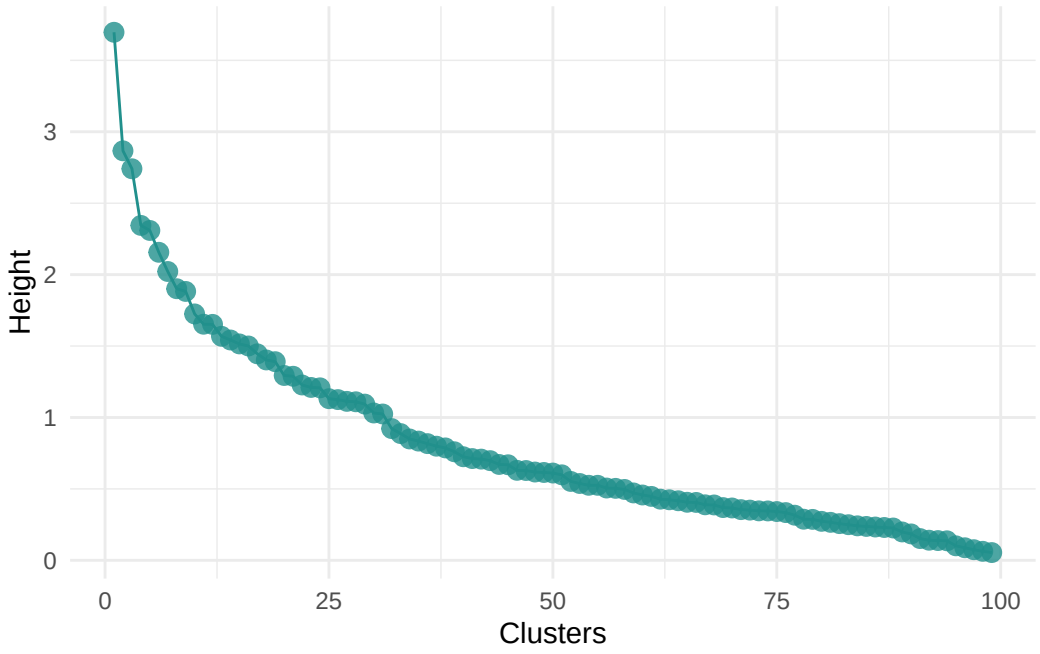


No método *average*, o dendograma também sugere 3 clusters:

```r
# criando learner
learner <- lrn(
  "clust.hclust",
  distmethod = "euclidean",
  method = "average"
)

# treinando o modelo
model <- learner$train(task)

# plotando dendograma
autoplot(model)
```




```r
# plotando altura x número de clusters
autoplot(model, type = "scree")
```



1.3 Anova

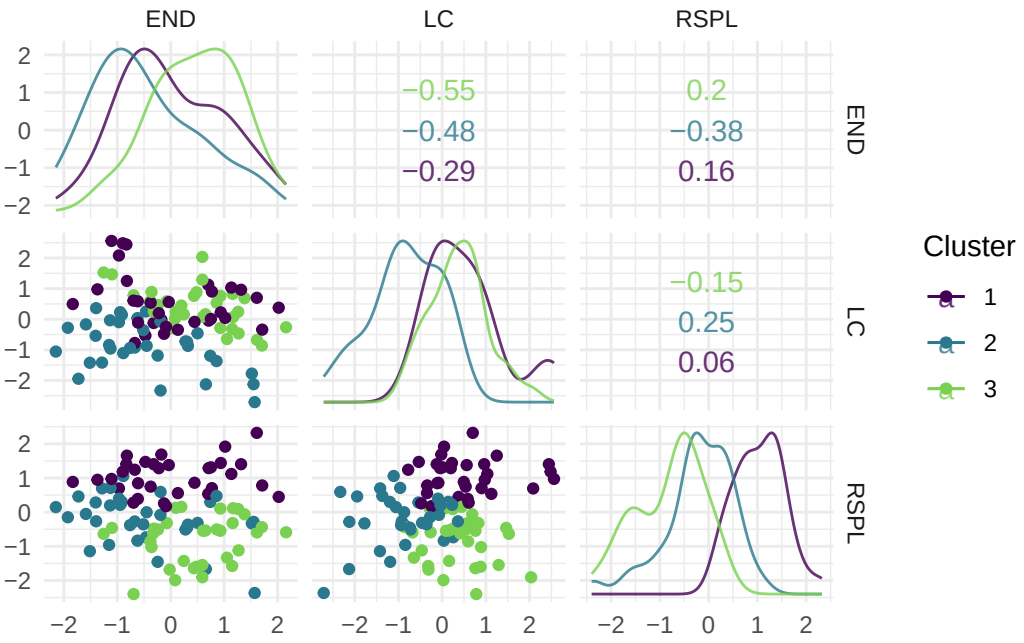
```
# criando learner
learner <- lrn(
  "clust.hclust",
  distmethod = "euclidean",
  method = "complete",
  k = 3
)

# treinando o modelo
model <- learner$train(task)

# predição
pred <- model$predict(task)
```

```
Warning: Learner 'clust.hclust' doesn't predict on new data and predictions may
not make sense on new data.
```

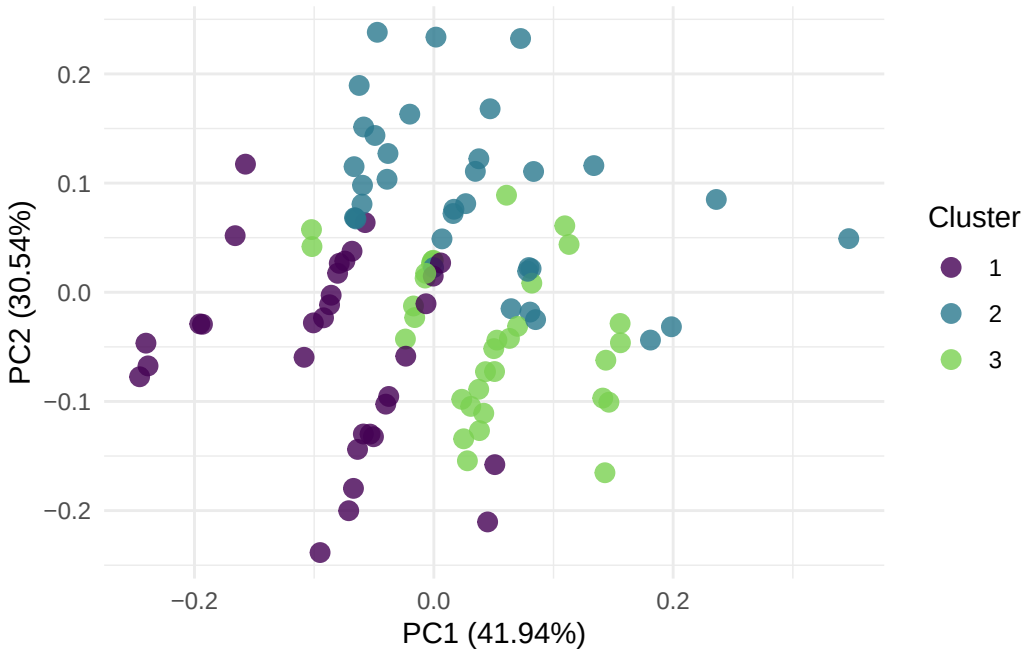
```
# visualizando os clusters
autoplot(pred, task)
```

```
Registered S3 method overwritten by 'GGally':
  method from
  +.gg   ggplot2
```

```
Warning in GGally::ggscatmat(data, color = "partition"): Factor variables are
omitted in plot
```




```
autoplot(pred, task, type = "pca")
```




```r
# bind com dados
dados_pred <- cbind(dados, as.data.table(pred))

# ANOVA para variável LC
aov(LC ~ partition, data = dados_pred) ▷ summary()
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
partition    1  0.051 0.05131   0.334  0.565
Residuals   98 15.057 0.15364
```

```r
# ANOVA para variável END
aov(END ~ partition, data = dados_pred) ▷ summary()
```

```
            Df Sum Sq  Mean Sq F value Pr(>F)
partition    1 0.0236 0.023574   3.467 0.0656 .
Residuals   98 0.6663 0.006799
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# ANOVA para variável RSPL
aov(RSPL ~ partition, data = dados_pred) ▷ summary()
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
partition    1 0.9804  0.9804   126.9 <2e-16 ***
Residuals   98 0.7570  0.0077
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

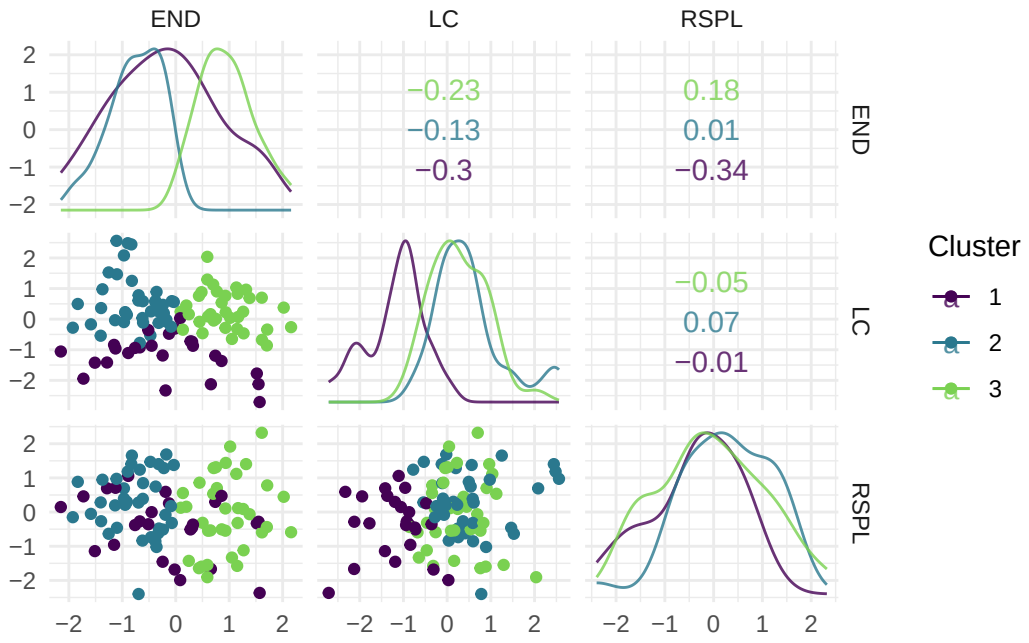
Para a variável LC e END, a ANOVA sugere que não há diferenças significativas entre os clusters. Já para a variável RSPL, há diferenças significativas entre os clusters.

1.4 Método de K-means

Utilizando agrupamento não hierarquico, podemos testar o método de K-means:

```r
# criando learner
learner <- lrn(
  "clust.kmeans",
  centers = 3
)

# treinando o modelo
model <- learner$train(task)

# predição
pred <- model$predict(task)

# visualizando os clusters
autoplot(pred, task)
```

```
Warning in GGally::ggscatmat(data, color = "partition"): Factor variables are
omitted in plot
```




```r
# bind com dados
dados_pred <- cbind(dados, as.data.table(pred))

# ANOVA para variável LC
aov(LC ~ partition, data = dados_pred) ▷ summary()
```

```
            Df Sum Sq Mean Sq F value   Pr(>F)
partition    1  3.991   3.991   35.19 4.51e-08 ***
Residuals   98 11.117   0.113
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

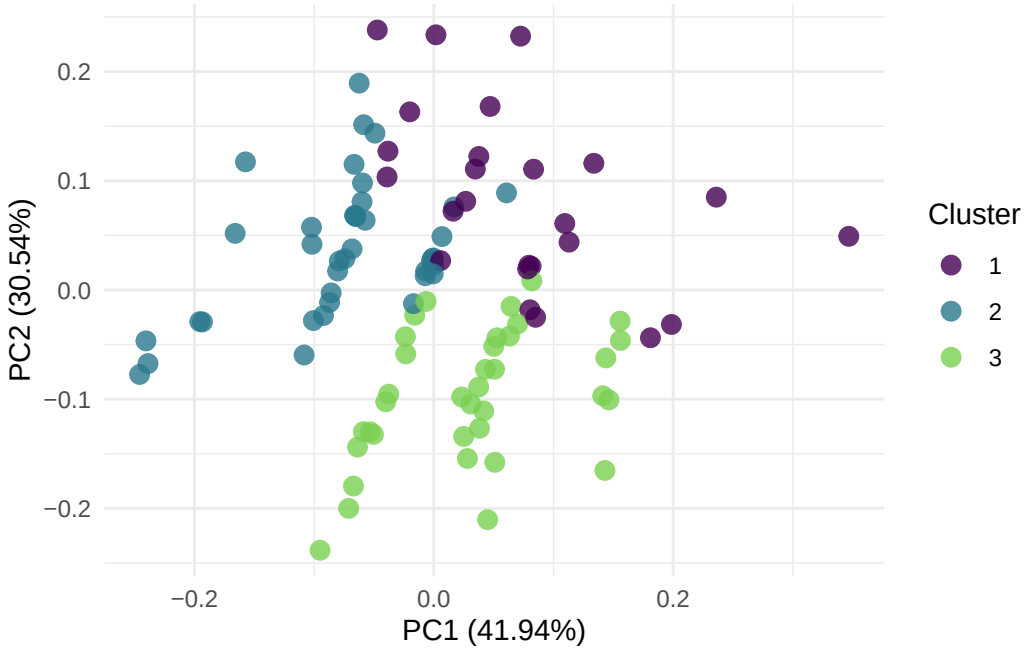
```r
# ANOVA para variável END
aov(END ~ partition, data = dados_pred) ▷ summary()
```

```
            Df Sum Sq Mean Sq F value   Pr(>F)
partition    1 0.1793 0.17933   34.42 6.03e-08 ***
Residuals   98 0.5106 0.00521
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# ANOVA para variável RSPL
aov(RSPL ~ partition, data = dados_pred) ▷ summary()
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
partition     1 0.0197 0.01973   1.125  0.291
Residuals    98 1.7177 0.01753
```

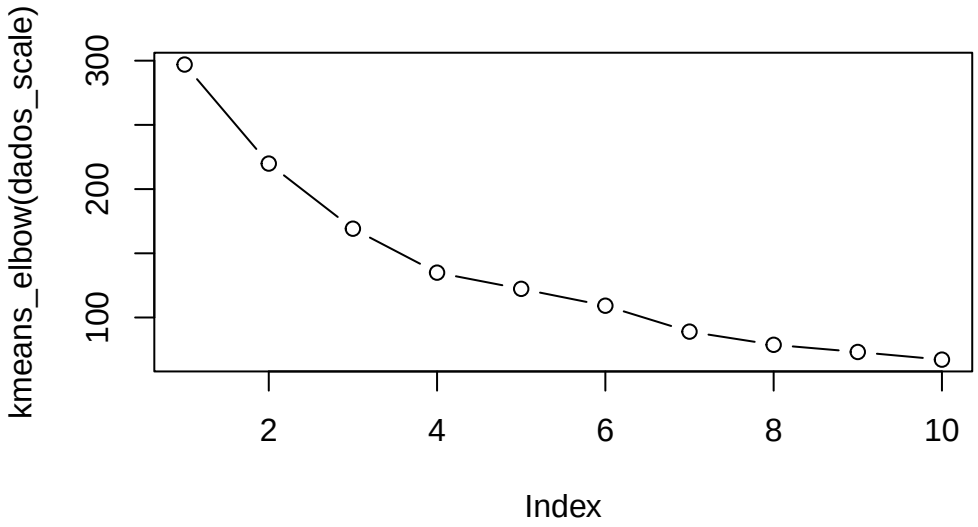
```r
# visualizando os clusters
autoplot(pred, task, type = "pca")
```



Usando método K-Means, a relação de significância da ANOVA se inverteu: para LC e END, há diferenças significativas entre os clusters, enquanto para RSPL, não há diferenças significativas.

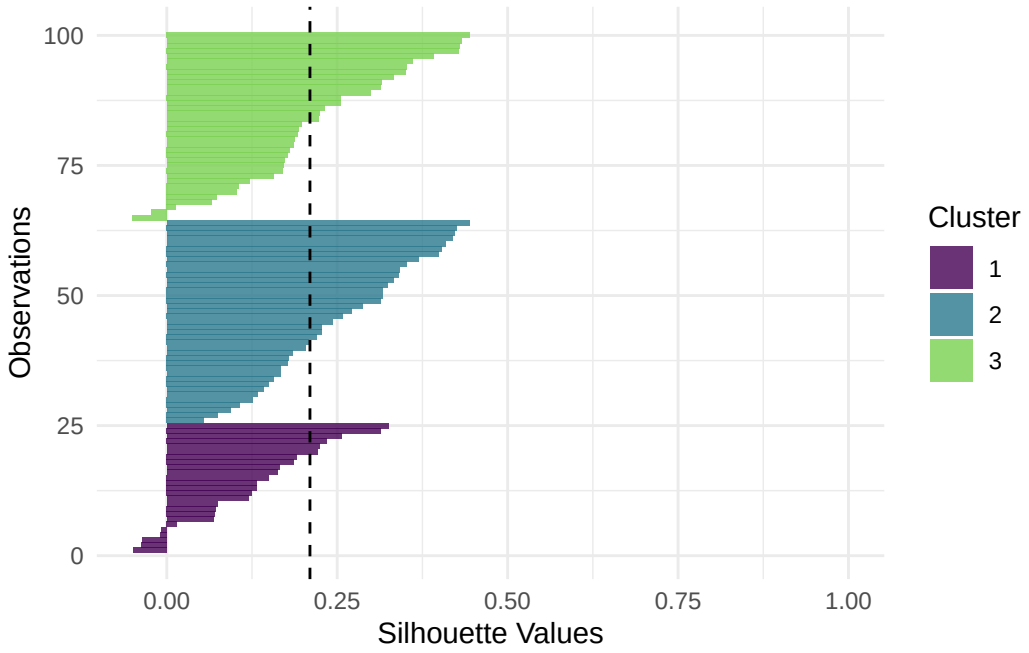
```r
# definição de k ótimo via elbow
kmeans_elbow <- function(data, k_max = 10) {
  wss <- numeric(k_max)
  for (k in 1:k_max) {
    model <- kmeans(data, centers = k)
    wss[k] <- model$tot.withinss
  }
  return(wss)
}

```

```r
# plotando elbow
dados_scale |> kmeans_elbow() |> plot(type = "b")
```



Por esse método, até 4 clusters poderiam ser considerados.

```r
# silhouette plot
autoplot(pred, task, type = "sil")
```



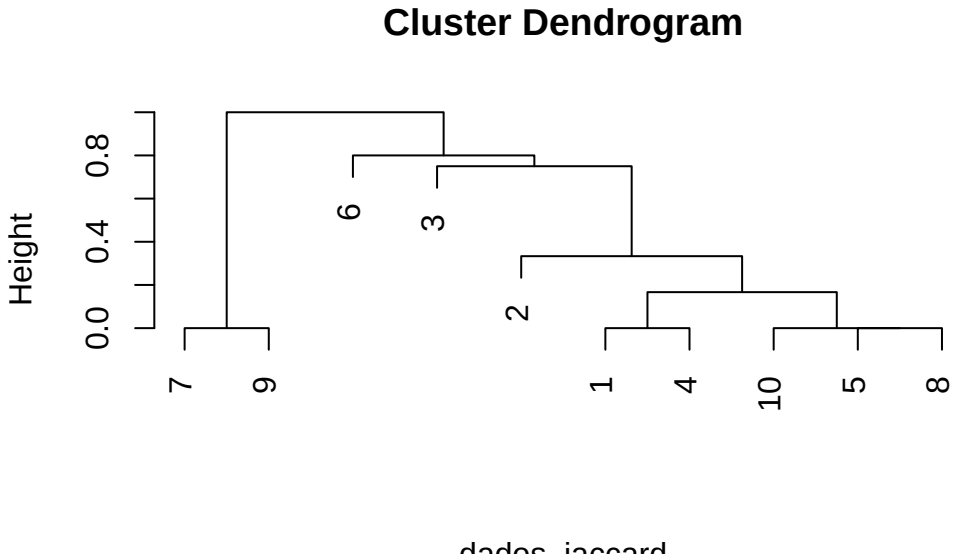
Poucas observações estão negativas, sugerindo que o método obteve separação aceitável.

2 Exercício 2

O dendograma sugere dois *clusters*: um com as empresas 7 e 9, e outro com as demais, uma vez que o *split* em 3 *clusters* não traz ganho de informação. E isso fica evidente observando os dados, já que as empresas 7 e 9 são iguais.

As observações 10, 5 e 8 também são iguais, mas para que formasem un *cluster*, seria necessário um *split* em 6 *clusters*, o que é demais num *dataset* de 10 observações.

```
# Carregando os dados
dados <- read.table("data-raw/clustering/empresas.txt", header = TRUE)

# corrigindo nomes
dados <- janitor::clean_names(dados)

# distância de Jaccard
dados_jaccard <- subset(dados, select = -empresa) ▷
  proxy::dist(method = "Jaccard")

# clusterização hierárquica
hc <- hclust(dados_jaccard, method = "single")

# plotando dendograma
plot(hc)
```



O primeiro grupo não realiza planejamento sucessório, então recomendaria fazê-lo. Já no segundo grupo, 75% realiza planejamento sucessório, evidenciando boa prática.