

## Tópico: Análise de Clusters no R

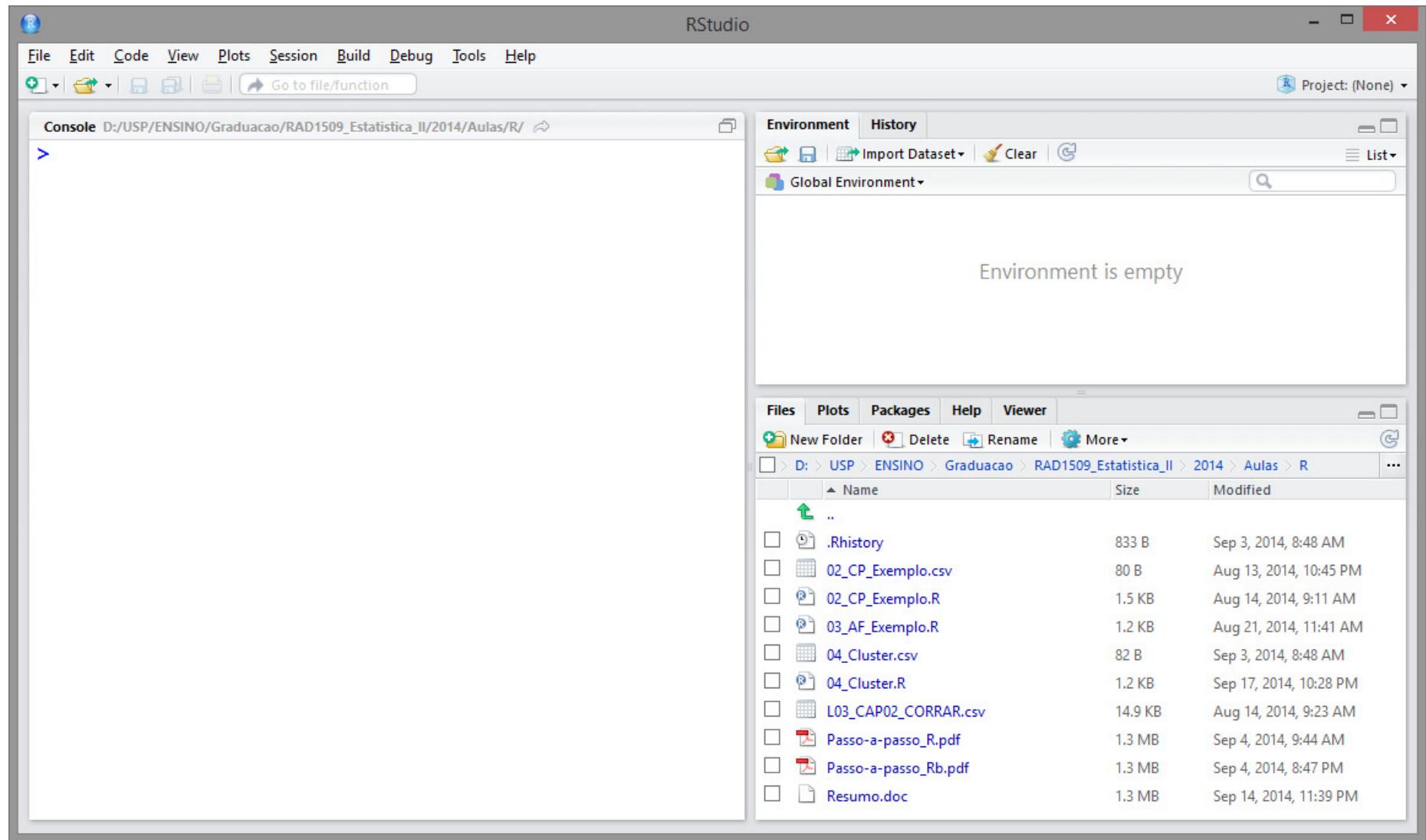
### Base de dados

**04\_Cluster.csv**

1	Empresa	gastos	lucro	
2	E1	50	70	
3	E2	35	45	
4	E3	32	45	
5	E4	52	60	
6	E5	30	40	
7	E6	45	48	
8				
9				
10				

# Análise de Agrupamentos no R

Considere o diretório de trabalho no qual esteja o arquivo .csv

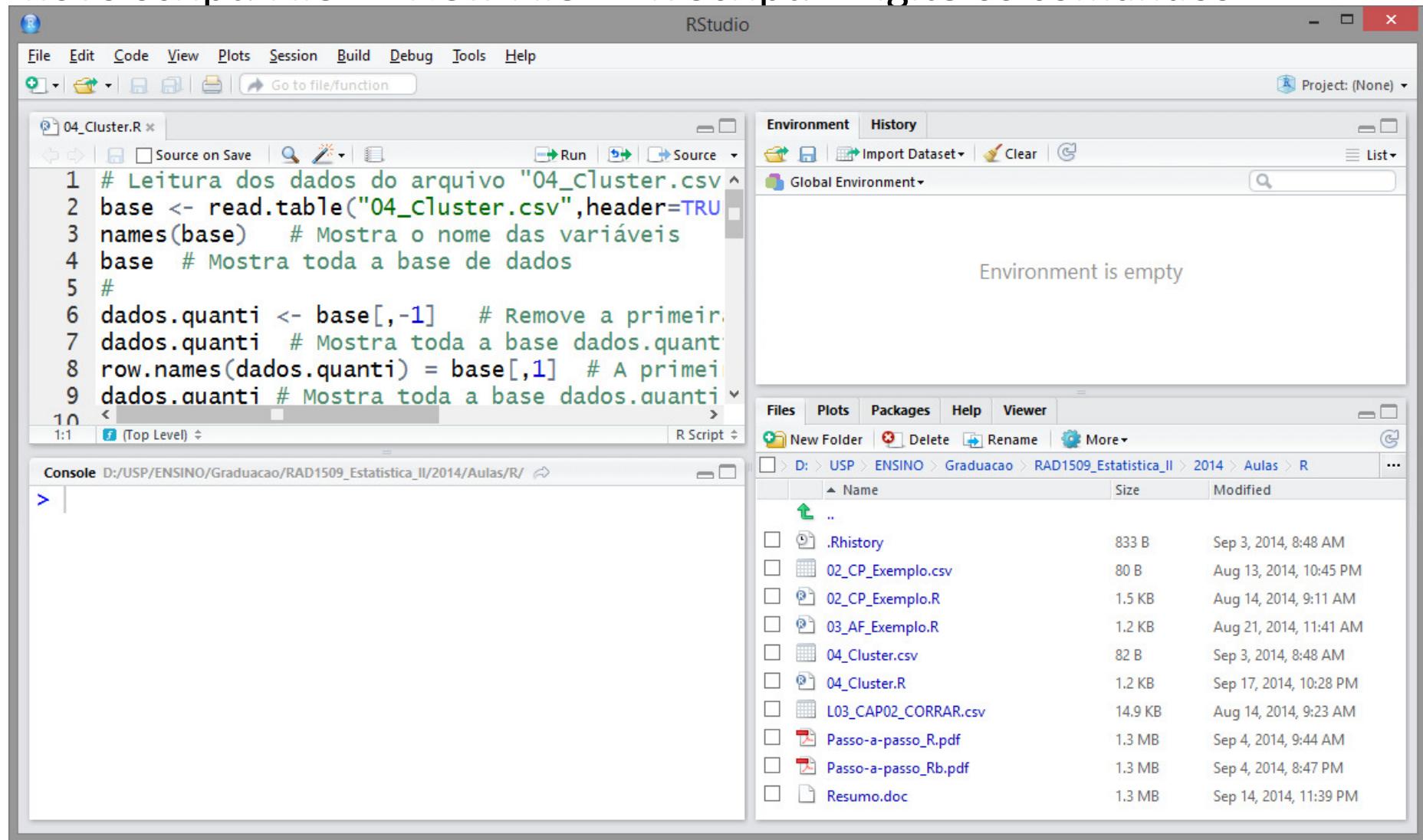


The screenshot shows the RStudio interface. The top menu bar includes File, Edit, Code, View, Plots, Session, Build, Debug, Tools, and Help. The left sidebar has tabs for Console, Files, Plots, Packages, Help, and Viewer, with 'Console' currently selected. The console pane shows the path D:/USP/ENSINO/Graduacao/RAD1509\_Estatistica\_II/2014/Aulas/R/. The environment pane displays the message 'Environment is empty'. The files pane shows the following directory structure and files:

Name	Size	Modified
..		
.Rhistory	833 B	Sep 3, 2014, 8:48 AM
02_CP_Exemplo.csv	80 B	Aug 13, 2014, 10:45 PM
02_CP_Exemplo.R	1.5 KB	Aug 14, 2014, 9:11 AM
03_AF_Exemplo.R	1.2 KB	Aug 21, 2014, 11:41 AM
04_Cluster.csv	82 B	Sep 3, 2014, 8:48 AM
04_Cluster.R	1.2 KB	Sep 17, 2014, 10:28 PM
L03_CAP02_CORRAR.csv	14.9 KB	Aug 14, 2014, 9:23 AM
Passo-a-passo_R.pdf	1.3 MB	Sep 4, 2014, 9:44 AM
Passo-a-passo_Rb.pdf	1.3 MB	Sep 4, 2014, 8:47 PM
Resumo.doc	1.3 MB	Sep 14, 2014, 11:39 PM

# Análise de Agrupamentos no R

Clique no arquivo “.R” para abrir o script, ou então comece escrever um novo script: File → New File → R Script. Digite os comandos



The screenshot shows the RStudio interface with the following components:

- Script Editor:** Displays the R script "04\_Cluster.R" with code for reading a CSV file and preparing it for clustering.
- Environment View:** Shows the message "Environment is empty".
- File Browser:** Shows a list of files in the directory D:/USP/ENSINO/Graduacao/RAD1509\_Estatistica\_II/2014/Aulas/R.

```
04_Cluster.R
1 # Leitura dos dados do arquivo "04_Cluster.csv"
2 base <- read.table("04_Cluster.csv", header=TRUE)
3 names(base) # Mostra o nome das variáveis
4 base # Mostra toda a base de dados
5 #
6 dados.quanti <- base[,-1] # Remove a primeira coluna
7 dados.quanti # Mostra toda a base dados.quanti
8 row.names(dados.quanti) = base[,1] # A primeira coluna é o nome da amostra
9 dados.quanti # Mostra toda a base dados.quanti
```

# Análise de Agrupamentos no R

Faça a leitura dos dados do arquivo .csv para o data.frame **base**.  
Observe o conteúdo do data.frame base.

The screenshot shows the RStudio interface with the following components:

- Code Editor (Top Left):** Displays the script `04_Cluster.R` containing R code for reading a CSV file and creating a data frame named `base`.
- Environment (Top Right):** Shows the `base` data frame with 6 observations and 3 variables.
- Console (Bottom Left):** Displays the output of running the script, including the names of the variables (`Empresa`, `gastos`, `lucro`) and the contents of the `base` data frame.
- Files (Bottom Right):** Shows a list of files in the current directory, including `04_Cluster.R` and other R scripts and documents.

```
04_Cluster.R
1 # Leitura dos dados do arquivo "04_cluster.csv"
2 base <- read.table("04_cluster.csv", header=TRUE)
3 names(base) # Mostra o nome das variáveis
4 base # Mostra toda a base de dados
5 #
6 dados.quanti <- base[,-1] # Remove a primeira coluna
7 dados.quanti # Mostra toda a base dados.quanti
8 row.names(dados.quanti) = base[,1] # A primeira coluna é o nome da empresa
9 dados.quanti # Mostra toda a base dados.quanti
```

```
Console D:/USP/ENSINO/Graduacao/RAD1509_Estatistica_II/2014/Aulas/R/
sep=";",dec=",")
> names(base) # Mostra o nome das variáveis
[1] "Empresa" "gastos" "lucro"
> base # Mostra toda a base de dados
  Empresa gastos lucro
1     E1     50    70
2     E2     35    45
3     E3     32    45
4     E4     52    60
5     E5     30    40
6     E6     45    48
>
```

Name	Size	Modified
.Rhistory	833 B	Sep 3, 2014, 8:48 AM
02_CP_Exemplo.csv	80 B	Aug 13, 2014, 10:45 PM
02_CP_Exemplo.R	1.5 KB	Aug 14, 2014, 9:11 AM
03_AF_Exemplo.R	1.2 KB	Aug 21, 2014, 11:41 AM
04_Cluster.csv	82 B	Sep 3, 2014, 8:48 AM
04_Cluster.R	1.2 KB	Sep 17, 2014, 10:28 PM
L03_CAP02_CORRAR.csv	14.9 KB	Aug 14, 2014, 9:23 AM
Passo-a-passo_R.pdf	1.3 MB	Sep 4, 2014, 9:44 AM
Passo-a-passo_Rb.pdf	1.3 MB	Sep 4, 2014, 8:47 PM
Resumo.doc	1.3 MB	Sep 14, 2014, 11:39 PM

# Análise de Agrupamentos no R

Remova a primeira coluna para ficar com dados quantitativos

The screenshot shows the RStudio interface with the following components:

- Code Editor:** Displays the script file `04_Cluster.R` containing R code to remove the first column from a dataset.
- Console:** Shows the output of the executed code, displaying the original dataset and the modified dataset `dados.quant` with the first column removed.
- Environment:** Shows the global environment with two objects: `base` (6 obs. of 3 variables) and `dados.quant` (6 obs. of 2 variables).
- Files:** Shows the file structure in the current directory: `D:/USP/ENSINO/Graduacao/RAD1509_Estatistica_II/2014/Aulas/R`.

Code in `04_Cluster.R`:

```
4 base # Mostra toda a base de dados
5 #
6 dados.quanti <- base[,-1] # Remove a primeira
7 dados.quanti # Mostra toda a base dados.quanti
8 row.names(dados.quanti) = base[,1] # A primeira
9 dados.quanti # Mostra toda a base dados.quanti
10 dados.quanti <- as.matrix(dados.quanti) # Considere
11 dados.quanti # Mostra toda a matriz dados.quanti
12 #
13 <--
```

Console output:

```
6 E6 45 48
> dados.quanti <- base[,-1] # Remove a primeira
coluna
> dados.quanti # Mostra toda a base dados.quanti
gastos lucro
1 50 70
2 35 45
3 32 45
4 52 60
5 30 40
6 45 48
>
```

File list in the Files panel:

Name	Size	Modified
.Rhistory	833 B	Sep 3, 2014, 8:48 AM
02_CP_Exemplo.csv	80 B	Aug 13, 2014, 10:45 PM
02_CP_Exemplo.R	1.5 KB	Aug 14, 2014, 9:11 AM
03_AF_Exemplo.R	1.2 KB	Aug 21, 2014, 11:41 AM
04_Cluster.csv	82 B	Sep 3, 2014, 8:48 AM
04_Cluster.R	1.2 KB	Sep 17, 2014, 10:28 PM
L03_CAP02_CORRAR.csv	14.9 KB	Aug 14, 2014, 9:23 AM
Passo-a-passo_R.pdf	1.3 MB	Sep 4, 2014, 9:44 AM
Passo-a-passo_Rb.pdf	1.3 MB	Sep 4, 2014, 8:47 PM
Resumo.doc	1.3 MB	Sep 14, 2014, 11:39 PM

# Análise de Agrupamentos no R

Coloque nomes nas linhas de acordo com os casos (1ª coluna da base)

The screenshot shows the RStudio interface with the following components:

- Script Editor (Top Left):** Displays the R script `04_Cluster.R` containing code to manipulate a dataset. The code includes comments like "# Mostra toda a base de dados" and "# Remove a primeira coluna da base contém os casos".
- Environment Browser (Top Right):** Shows the global environment with two objects:
  - `base`: 6 obs. of 3 variables
  - `dados.quant...`: 6 obs. of 2 variables
- Console (Bottom Left):** Displays the output of the R script, including the first few rows of the dataset and the results of the clustering command.
- File Browser (Bottom Right):** Shows the file structure at `D:/USP/ENSINO/Graduacao/RAD1509_Estatistica_II/2014/Aulas/R`, listing files like `.Rhistory`, `02_CP_Exemplo.csv`, and `04_Cluster.R`.

# Análise de Agrupamentos no R

Mude a base dados.quanti para uma forma de matriz para poder obter a distância entre os casos.

The screenshot shows the RStudio interface with the following components:

- Script Editor:** The "04\_Cluster.R" script contains R code for clustering. The last few lines of the code are highlighted:

```
8 row.names(dados.quanti) = base[,1] # A primeira coluna é o nome das empresas
9 dados.quanti # Mostra toda a base dados.quanti
10 dados.quanti <- as.matrix(dados.quanti) # Considera dados.quanti como matriz
11 dados.quanti # Mostra toda a matriz dados.quanti
12 #
13 # Cálculo de distâncias entre os casos
14 d <- dist(dados.quanti, method = "euclidean")
15 d
16 # Considerando a distância euclideana ao quadrado
17 <--
```
- Console:** Displays the output of the R code. It shows the data structure and the results of the clustering analysis:

```
E6      45     48
> dados.quanti <- as.matrix(dados.quanti) # Considera dados.quanti como matriz
> dados.quanti # Mostra toda a matriz dados.quanti
   gastos lucro
E1      50     70
E2      35     45
E3      32     45
E4      52     60
E5      30     40
E6      45     48
>
```
- Environment:** Shows the global environment with variables "base" and "dados.quanti".

Name	Type	Value
base	base	6 obs. of 3 variables
dados.quanti	matrix	int [1:6, 1:2] 50 35 32 52 30 ...
- File Browser:** Shows the file structure in the current directory:

Name	Size	Modified
.Rhistory	833 B	Sep 3, 2014, 8:48 AM
02_CP_Exemplo.csv	80 B	Aug 13, 2014, 10:45 PM
02_CP_Exemplo.R	1.5 KB	Aug 14, 2014, 9:11 AM
03_AF_Exemplo.R	1.2 KB	Aug 21, 2014, 11:41 AM
04_Cluster.csv	82 B	Sep 3, 2014, 8:48 AM
04_Cluster.R	1.2 KB	Sep 17, 2014, 10:28 PM
L03_CAP02_CORRAR.csv	14.9 KB	Aug 14, 2014, 9:23 AM
Passo-a-passo_R.pdf	1.3 MB	Sep 4, 2014, 9:44 AM
Passo-a-passo_Rb.pdf	1.3 MB	Sep 4, 2014, 8:47 PM
Resumo.doc	1.3 MB	Sep 14, 2014, 11:39 PM

# Análise de Agrupamentos no R

Calcule a matriz de distâncias entre os casos E1, E2, ...

The screenshot shows the RStudio interface with the following components:

- Code Editor:** Displays the script file `04_Cluster.R*` containing R code for distance calculation and hierarchical clustering.
- Console:** Shows the output of the R code, including the distance matrix between cases E1 through E6.
- Environment:** Shows the global environment with variables `base` and `d`.
- Files:** Shows the project directory structure and files listed in the sidebar.

**R Code in `04_Cluster.R*`:**

```
11 dados.quanti # Mostra toda a matriz dados.quanti
12 #
13 # Cálculo de distâncias entre os casos
14 d <- dist(dados.quanti, method = "euclidean")
15 d
16 # Considerando a distância euclidiana ao quadrado:
17 d2 <- d^2
18 d2
19 # Cluster Hierárquico, método:
20 <--
```

**Console Output:**

```
E5      30     40
E6      45     48
> # Cálculo de distâncias entre os casos
> d <- dist(dados.quanti, method = "euclidean")
> d
      E1      E2      E3      E4      E5
E2 29.154759
E3 30.805844 3.000000
E4 10.198039 22.671568 25.000000
E5 36.055513 7.071068 5.385165 29.732137
E6 22.561028 10.440307 13.341664 13.892444 17.000000
>
```

**Project Files:**

- `.Rhistory` (833 B, Sep 3, 2014, 8:48 AM)
- `02_CP_Exemplo.csv` (80 B, Aug 13, 2014, 10:45 PM)
- `02_CP_Exemplo.R` (1.5 KB, Aug 14, 2014, 9:11 AM)
- `03_AF_Exemplo.R` (1.2 KB, Aug 21, 2014, 11:41 AM)
- `04_Cluster.csv` (82 B, Sep 3, 2014, 8:48 AM)
- `04_Cluster.R` (1.2 KB, Sep 17, 2014, 10:28 PM)
- `L03_CAP02_CORRAR.csv` (14.9 KB, Aug 14, 2014, 9:23 AM)
- `Passo-a-passo_R.pdf` (1.3 MB, Sep 4, 2014, 9:44 AM)
- `Passo-a-passo_Rb.pdf` (1.3 MB, Sep 4, 2014, 8:47 PM)
- `Resumo.doc` (1.3 MB, Sep 14, 2014, 11:39 PM)

# Análise de Agrupamentos no R

Calcule a distância euclidiana ao quadrado.

The screenshot shows the RStudio interface with the following components:

- Code Editor:** Displays the script file `04_Cluster.R*` containing R code for data manipulation and clustering.
- Console:** Shows the output of the R code, including numerical data and the results of the clustering analysis.
- Environment:** Shows the global environment with objects `d` and `d2`.
- Data View:** Shows the data frame `base` with 6 observations and 3 variables.
- Files View:** Shows the file structure and list of files in the current directory.

**R Code in `04_Cluster.R`:**

```
11 dados.quanti # Mostra toda a matriz dados.quanti
12 #
13 # Cálculo de distâncias entre os casos
14 d <- dist(dados.quanti, method = "euclidean")
15 d
16 # Considerando a distância euclidiana ao quadrado:
17 d2 <- d^2
18 d2
19 # Cluster Hierárquico, método:
20 <
```

**Console Output:**

```
E5 36.055513 7.071068 5.385165 29.732137
E6 22.561028 10.440307 13.341664 13.892444 17.000000
> # Considerando a distância euclidiana ao quadrado:
> d2 <- d^2
> d2
    E1   E2   E3   E4   E5
E2  850
E3   949    9
E4   104   514   625
E5  1300    50    29   884
E6   509   109   178   193   289
> |
```

**Environment View:**

Object	Type	Value
<code>d</code>	Class 'dist' atomic	[1:15] 2...
<code>d2</code>	Class 'dist' atomic	[1:15] 8...

**Data View:**

Variable	Value
<code>base</code>	6 obs. of 3 variables dados.quan.int [1:6, 1:2] 50 35 32 52...

**Files View:**

Name	Size	Modified
.Rhistory	833 B	Sep 3, 2014, 8:48 AM
02_CP_Exemplo.csv	80 B	Aug 13, 2014, 10:45 PM
02_CP_Exemplo.R	1.5 KB	Aug 14, 2014, 9:11 AM
03_AF_Exemplo.R	1.2 KB	Aug 21, 2014, 11:41 AM
04_Cluster.csv	82 B	Sep 3, 2014, 8:48 AM
04_Cluster.R	1.2 KB	Sep 17, 2014, 10:28 PM
L03_CAP02_CORRAR.csv	14.9 KB	Aug 14, 2014, 9:23 AM
Passo-a-passo_R.pdf	1.3 MB	Sep 4, 2014, 9:44 AM
Passo-a-passo_Rb.pdf	1.3 MB	Sep 4, 2014, 8:47 PM
Resumo.doc	1.3 MB	Sep 14, 2014, 11:39 PM

# Análise de Agrupamentos no R

Obtenção de Clusters pelo método hierárquico com vários métodos de ligação: simples, completa, média e Ward.

The screenshot shows the RStudio interface with the following components:

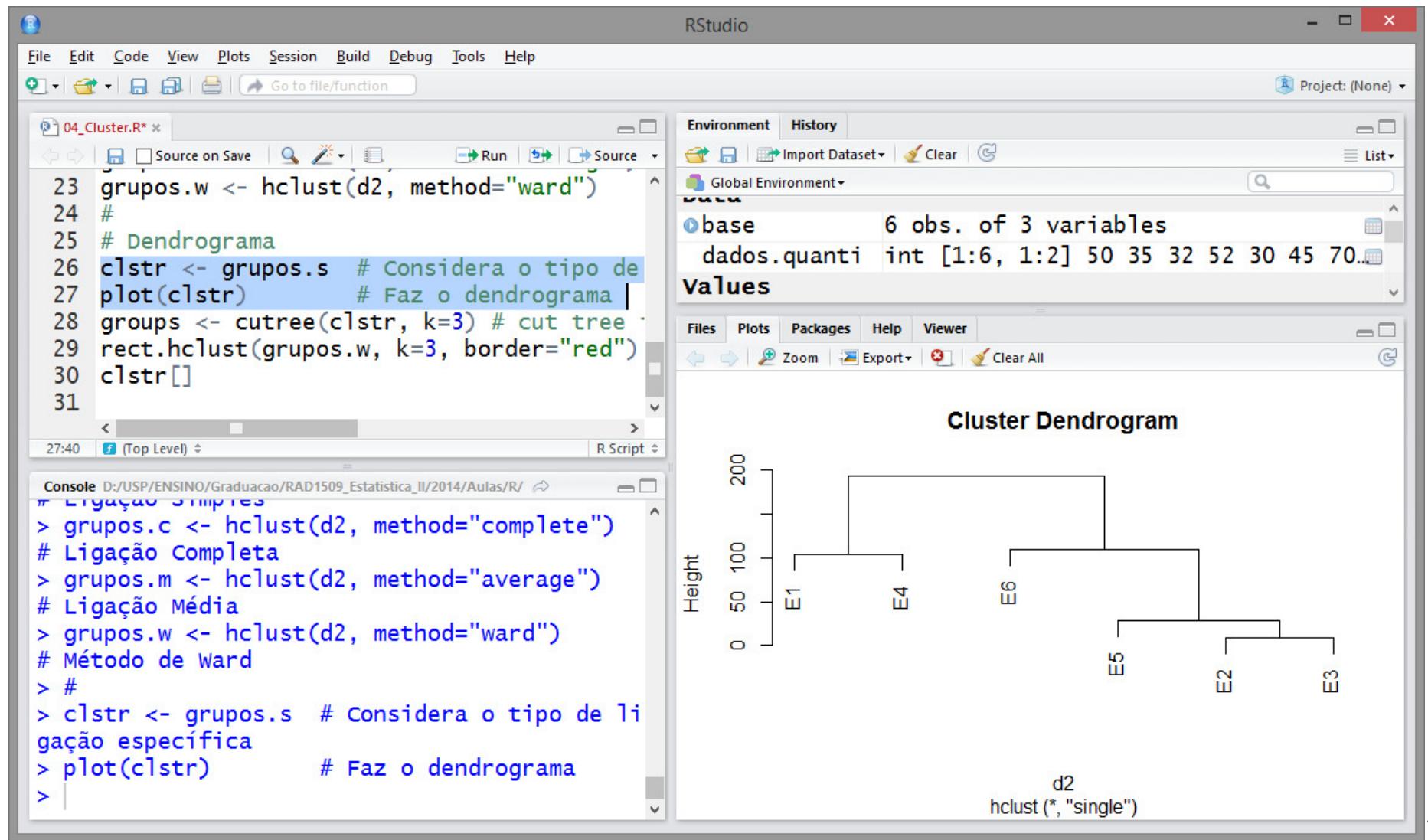
- Code Editor:** Displays the R script `04_Cluster.R*` containing the following code:

```
18 d2
19 # Cluster Hierárquico, método:
20 grupos.s <- hclust(d2, method="single")      # Ligação Simples
21 grupos.c <- hclust(d2, method="complete")       # Ligação Completa
22 grupos.m <- hclust(d2, method="average")        # Ligação Média
23 grupos.w <- hclust(d2, method="ward")           # Método de Ward
24 #
25 # Dendrograma
26 clstr <- arubos.s # Considera o tipo de ligação específica
```
- Console:** Shows the output of the R script:

```
E2 850
E3 949   9
E4 104  514  625
E5 1300  50   29   884
E6 509   109  178  193  289
> # Cluster Hierárquico, método:
> grupos.s <- hclust(d2, method="single")      # Ligação Simples
> grupos.c <- hclust(d2, method="complete")       # Ligação Completa
> grupos.m <- hclust(d2, method="average")        # Ligação Média
> grupos.w <- hclust(d2, method="ward")           # Método de Ward
> #
```
- Environment:** Shows the global environment variables `d` and `d2`.
- Data:** Shows the dataset `base` with 6 observations and 3 variables.
- Values:** Shows the types of `d` and `d2` as `dist` atomic vectors.
- File Explorer:** Shows the project structure with files like `.Rhistory`, `02_CP_Exemplo.csv`, `02_CP_Exemplo.R`, etc.

# Análise de Agrupamentos no R

Selecione um dos modelos feitos anteriormente e faça o dendrograma.



# Análise de Agrupamentos no R

Selecione um número de grupos adequado para ser observado no dendrograma.

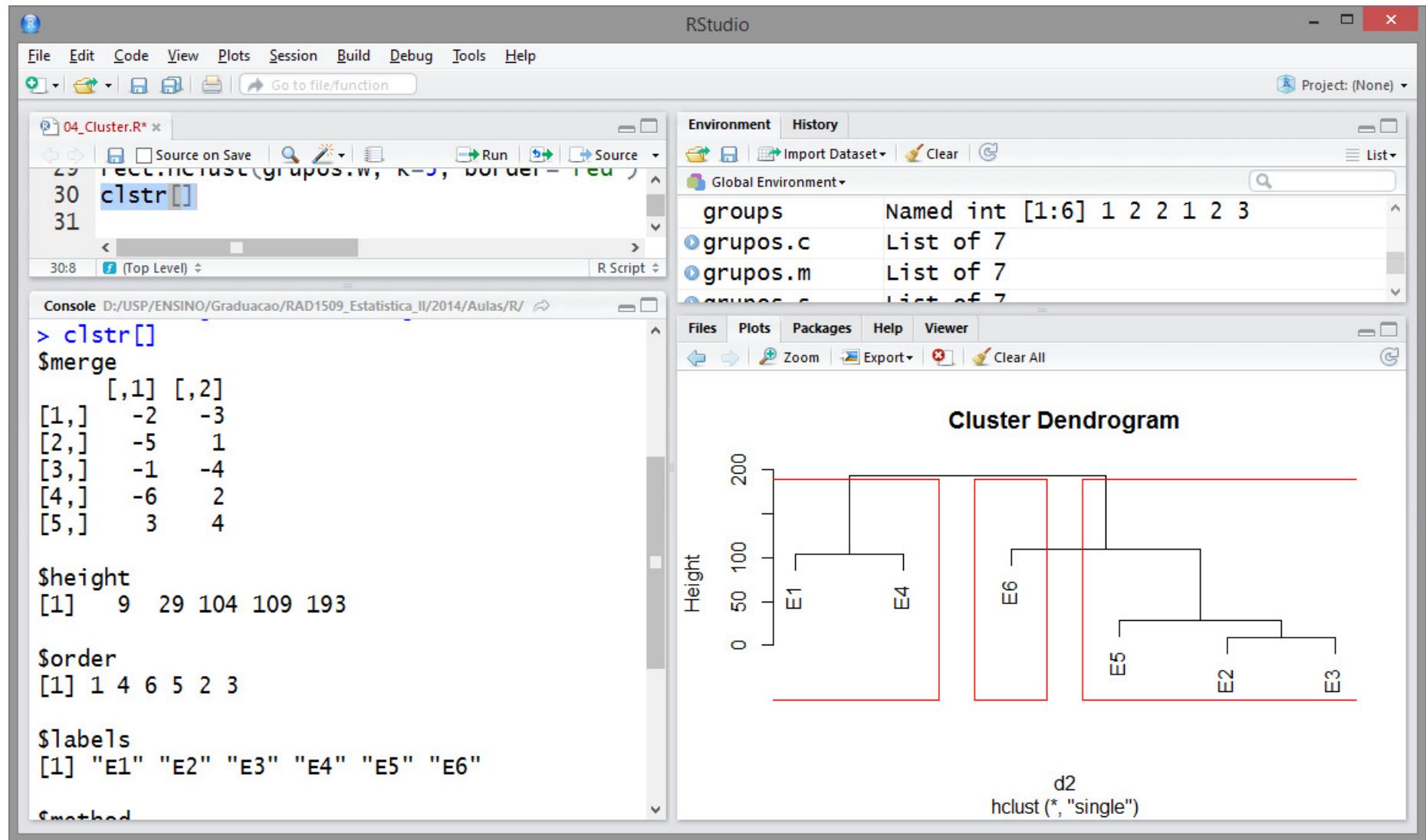
The screenshot shows the RStudio interface with the following components:

- Code Editor:** Displays the R script "04\_Cluster.R" containing code for clustering and plotting. Lines 28 and 29 are highlighted in blue: 

```
groups <- cutree(clstr, k=3) # cut tree into 3 clusters
rect.hclust(grupos.w, k=3, border="red") # insere retângulos no dendrograma
```
- Console:** Displays the R command history, including the execution of the highlighted code and its output.
- Environment View:** Shows the global environment with objects like "groups", "grupos.c", "grupos.m", and "grupos.s".
- Dendrogram Plot:** A "Cluster Dendrogram" plot titled "d2 hclust (\*, "single")". The plot shows six data points labeled E1 through E6 being grouped into three clusters. Red rectangles are drawn around each cluster to highlight the resulting groups.

# Análise de Agrupamentos no R

Verifique a análise completa.



# Análise de Agrupamentos no R

Verifique a análise completa.

The screenshot shows the RStudio interface with the following components:

- File menu:** File, Edit, Code, View, Plots, Session, Build, Debug, Tools, Help.
- Toolbar:** Source on Save, Run, Source, Import Dataset, Clear, List.
- Script Editor:** Shows the R script `04_Cluster.R*` with code for clustering six data points (E1-E6) using the `hclust` function.
- Console:** Displays the results of the clustering process.
- Environment:** Shows the variable `groups` as a named vector [1:6] with values 1, 2, 2, 1, 2, 3.
- Plots:** A "Cluster Dendrogram" plot showing the hierarchical clustering of six data points (E1-E6). The y-axis is labeled "Height" with ticks at 0, 50, 100, and 200. The x-axis is labeled "d2" with labels for each point: E1, E4, E6, E5, E2, E3. Red vertical lines group the points into three clusters: {E1, E4}, {E6}, and {E5, E2, E3}.
- Help:** Help, Viewer.
- Bottom status bar:** Shows the command `hclust (*, "single")`.

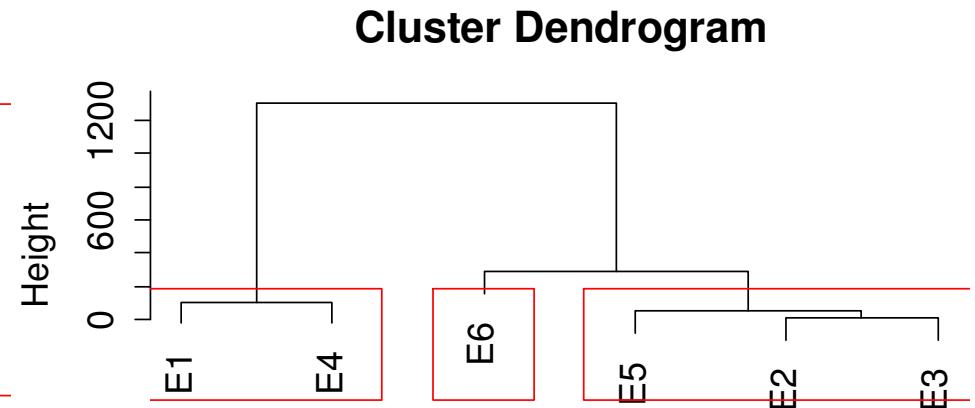
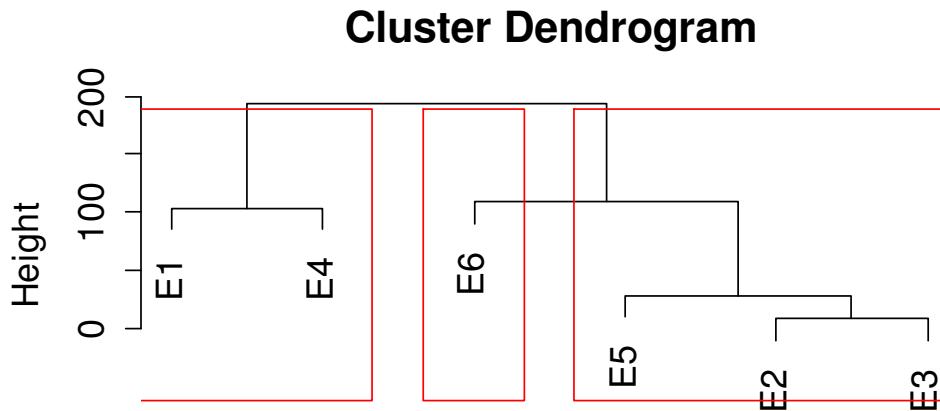
# Análise de Agrupamentos no R

Obtenha a lista dos membros de cada agrupamento

The screenshot shows the RStudio interface with the following components:

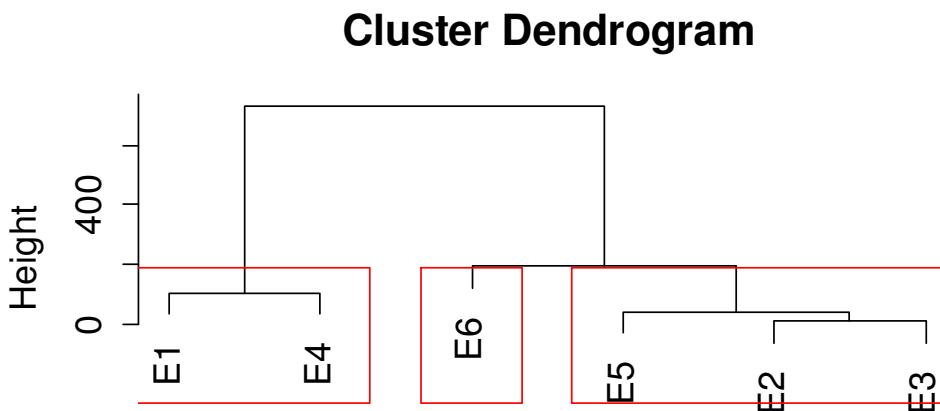
- Code Editor:** Displays the script `04_Cluster.R*` containing R code for hierarchical clustering.
- Console:** Shows the output of the command `> groups`, which prints the cluster assignments for six data points (E1-E6) into three groups: E1, E2, and E3.
- Environment View:** Shows the variable `groups` as a `Named int [1:6]` vector with values `1 2 2 1 2 3`.
- Plots View:** Displays a **Cluster Dendrogram** plot. The y-axis is labeled "Height" with ticks at 0, 50, 100, and 200. The x-axis is labeled "d2" and "hclust (\*, "single")". The plot shows a hierarchical clustering structure where data points E1, E2, E3, E4, and E5 form one group, while E6 forms a separate group. Red vertical lines indicate the assignment of each point to its respective cluster at the final step of the clustering process.

# Análise de Agrupamentos no R

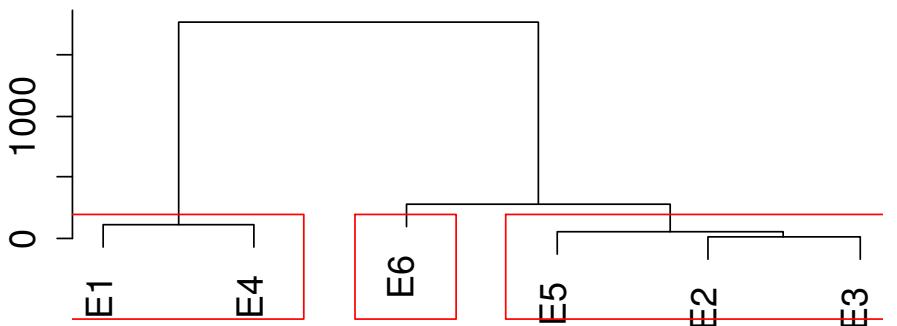


`d2  
hclust (*, "single")`

`d2  
hclust (*, "complete")`



`d2  
hclust (*, "average")`



`d2  
hclust (*, "ward")`