

Tópico: Análise de Clusters no R

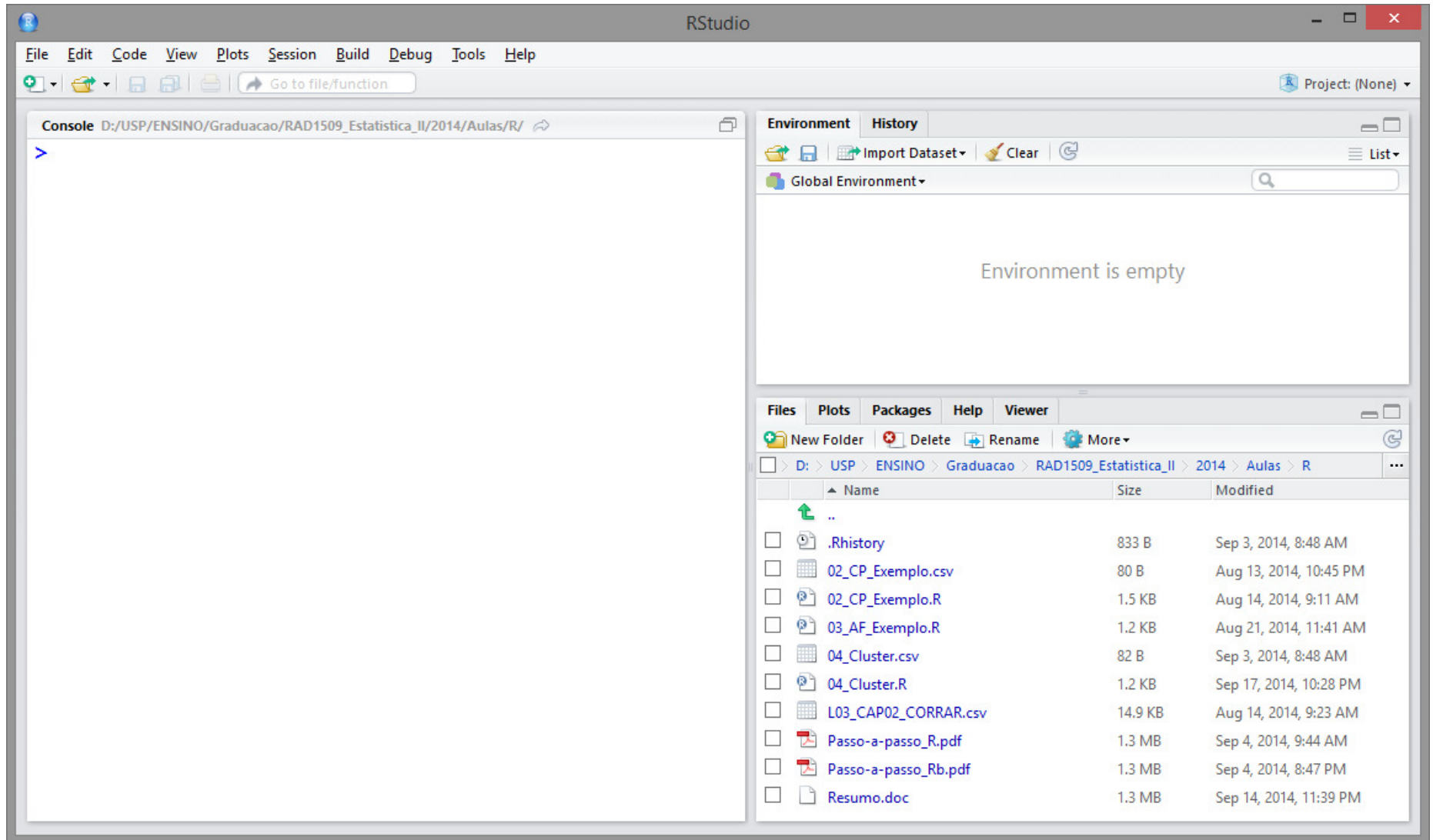
Base de dados

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**



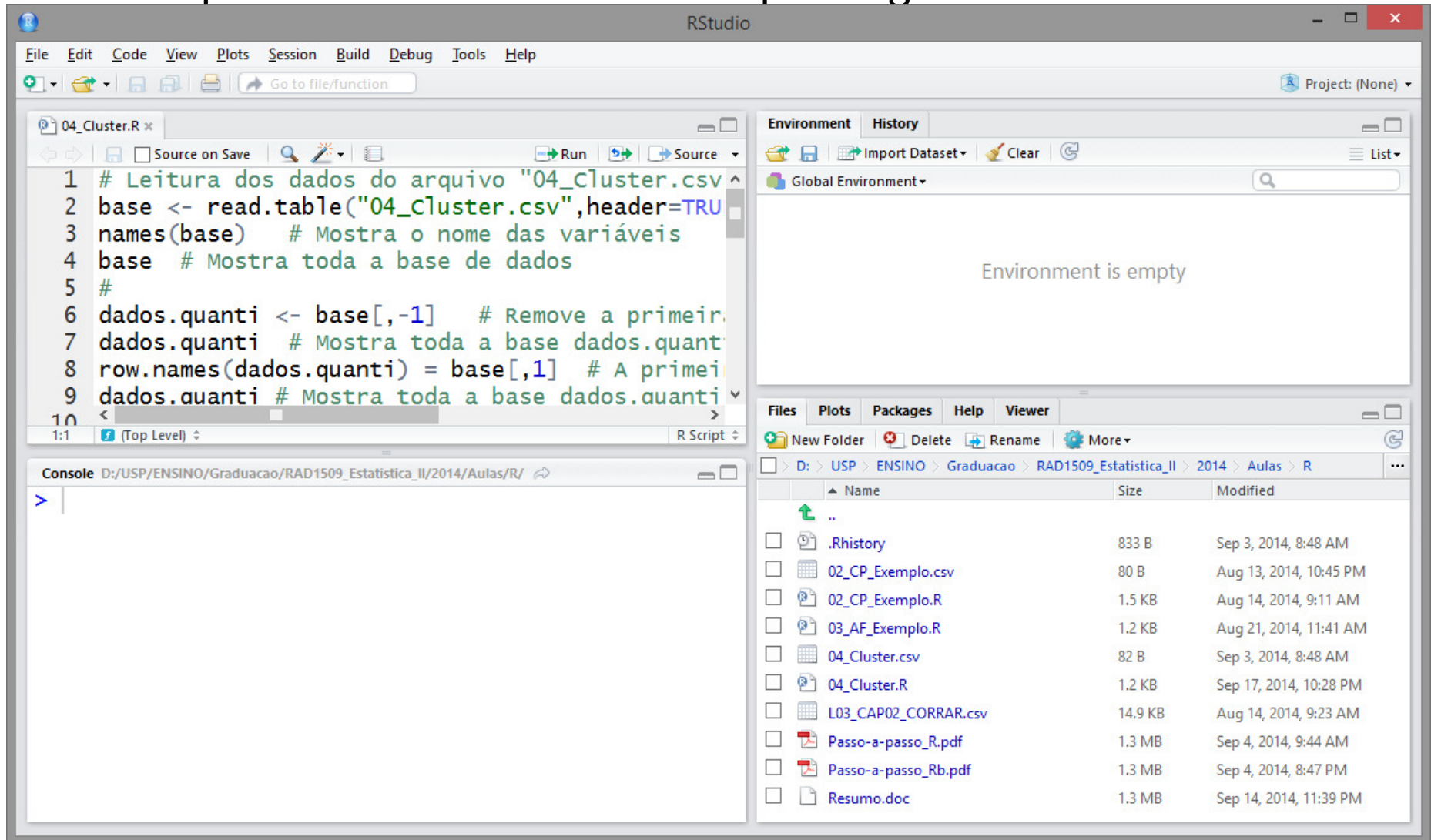
The screenshot shows the RStudio interface with the following components:

- Console:** Shows the current working directory as `D:/USP/ENSINO/Graduacao/RAD1509_Estatistica_II/2014/Aulas/R/`.
- Environment:** Shows the Global Environment, which is empty.
- Files:** Shows a list of files 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

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 environment with the following components:

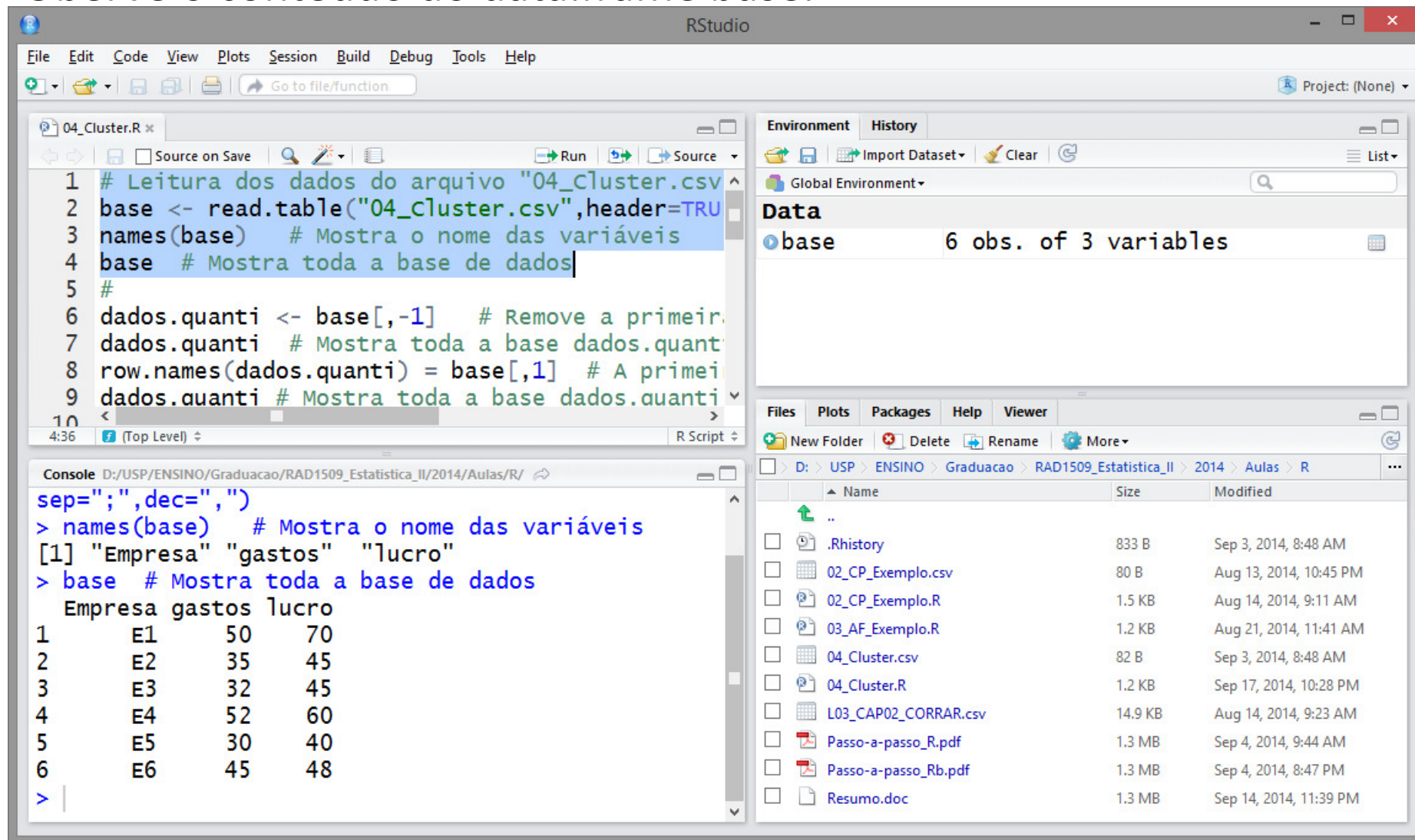
- Script Editor:** Contains the following R code:


```
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.quantil <- base[,-1] # Remove a primeira coluna
7 dados.quantil # Mostra toda a base dados.quantil
8 row.names(dados.quantil) = base[,1] # A primeira coluna como nomes de linha
9 dados.quantil # Mostra toda a base dados.quantil
```
- Environment Pane:** Shows "Global Environment" and "Environment is empty".
- File Explorer:** Displays the file structure of the project:

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

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:

- Source Editor:** Contains the R script for reading and processing the data.
- Environment:** Shows the 'base' data frame with 6 observations and 3 variables.
- Console:** Displays the output of the R commands, showing the names of the variables and the data content.
- Files Panel:** Lists the files in the current directory, including the CSV file '04_Cluster.csv'.

R Script (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.quantit <- base[,-1] # Remove a primeira coluna
7 dados.quantit # Mostra toda a base dados.quantit
8 row.names(dados.quantit) = base[,1] # A primeira coluna como nomes de linha
9 dados.quantit # Mostra toda a base dados.quantit
```

Console Output:

```
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
```

Environment:

Data

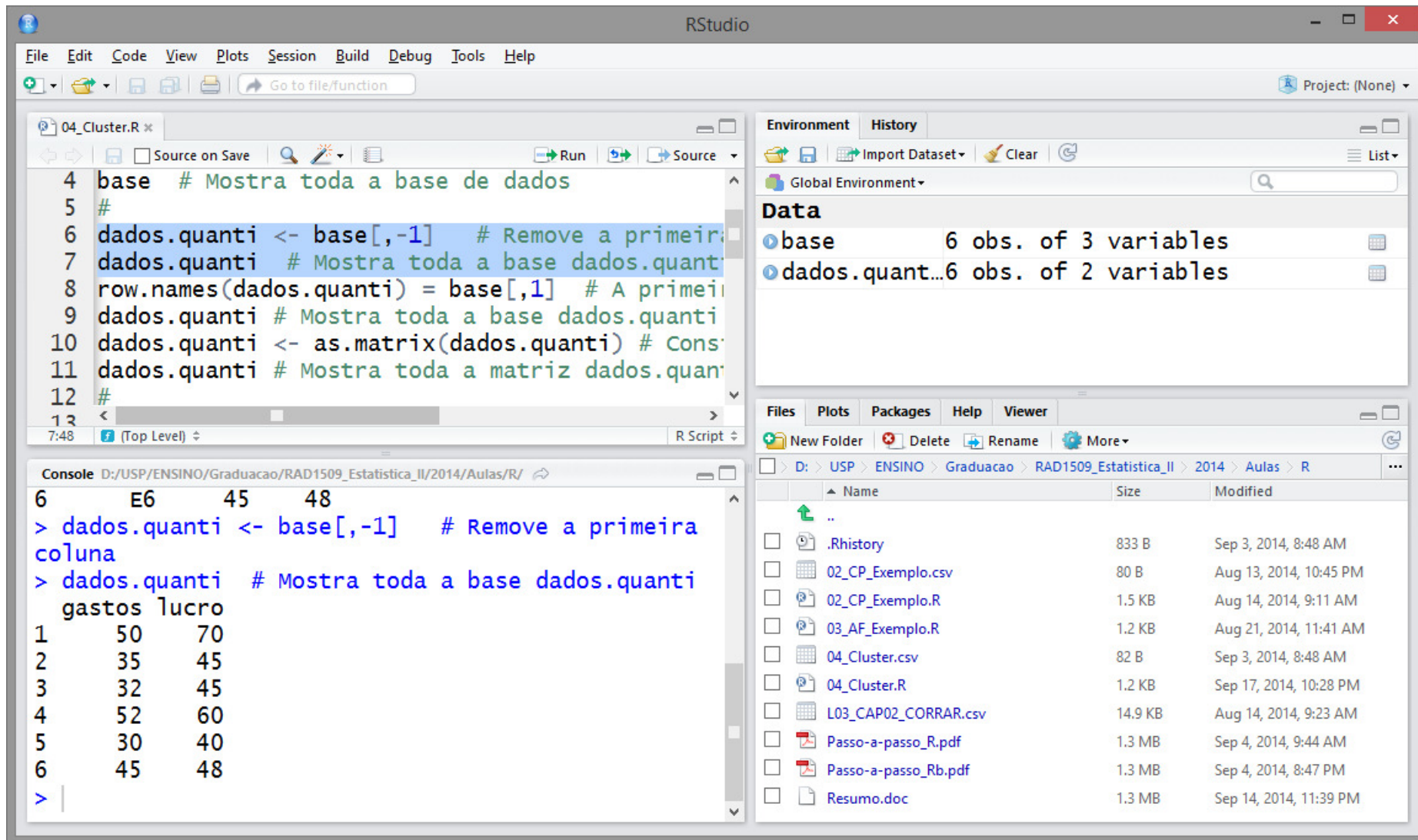
base 6 obs. of 3 variables

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

Remova a primeira coluna para ficar com dados quantitativos



The screenshot shows the RStudio interface with the following components:

- Source Editor:** Contains R code for removing the first column from a dataset.


```

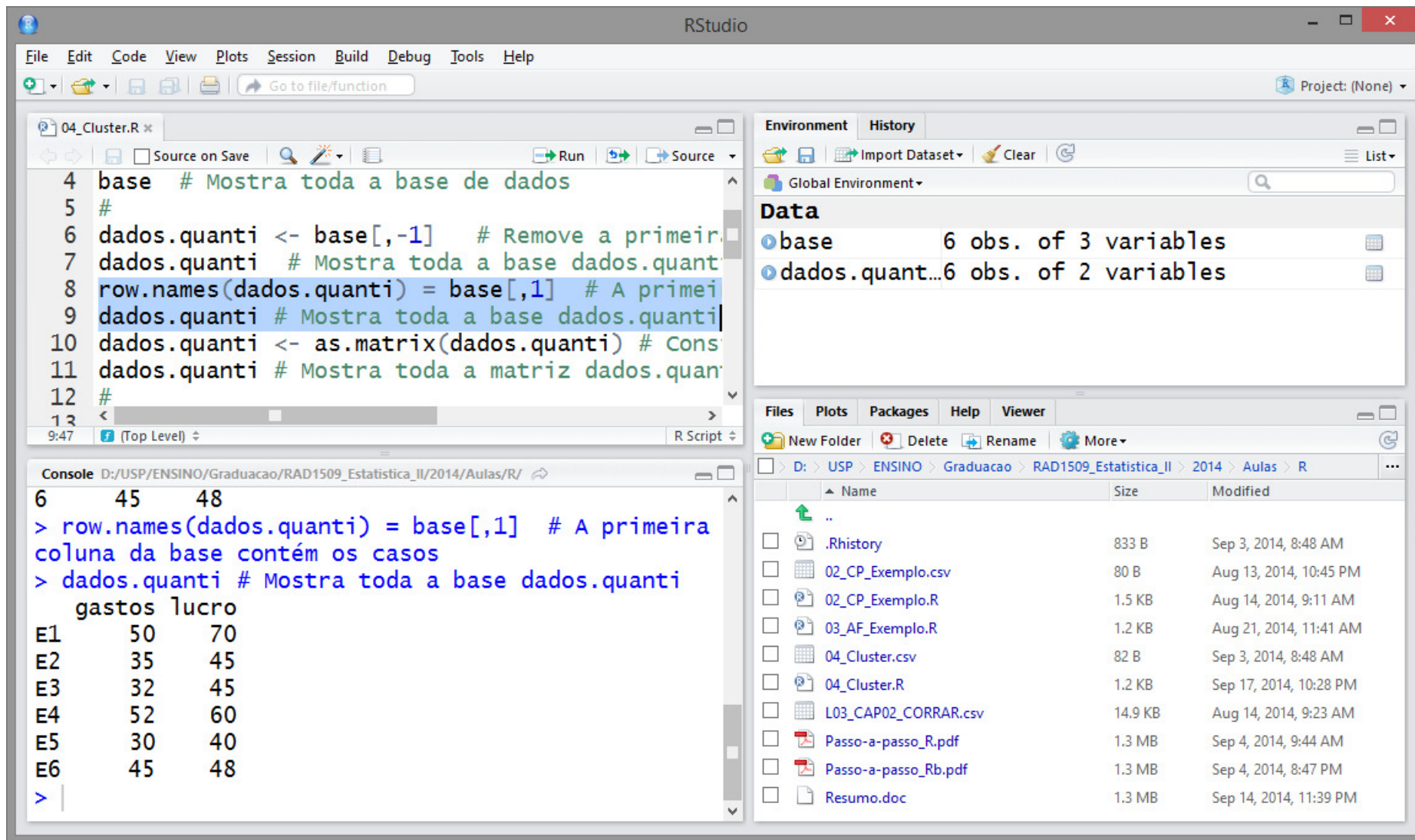
4 base # Mostra toda a base de dados
5 #
6 dados.quanti <- base[,-1] # Remove a primeira
7 dados.quanti # Mostra toda a base dados.quanti
8 row.names(dados.quanti) = base[,1] # A primeira
9 dados.quanti # Mostra toda a base dados.quanti
10 dados.quanti <- as.matrix(dados.quanti) # Converte
11 dados.quanti # Mostra toda a matriz dados.quanti
12 #
13 <
      
```
- Environment:** Shows the global environment with two data objects:
 - base:** 6 obs. of 3 variables
 - dados.quanti:** 6 obs. of 2 variables
- Console:** Shows the execution of the code, resulting in a matrix of quantitative data.


```

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
      
```
- Files:** Shows the project directory structure, including files like .Rhistory, 02_CP_Exemplo.csv, 02_CP_Exemplo.R, 03_AF_Exemplo.R, 04_Cluster.csv, 04_Cluster.R, L03_CAP02_CORRAR.csv, Passo-a-passo_R.pdf, Passo-a-passo_Rb.pdf, and Resumo.doc.

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:

- Source Editor:** Contains R code for data processing. Lines 8-9 are highlighted in blue.


```

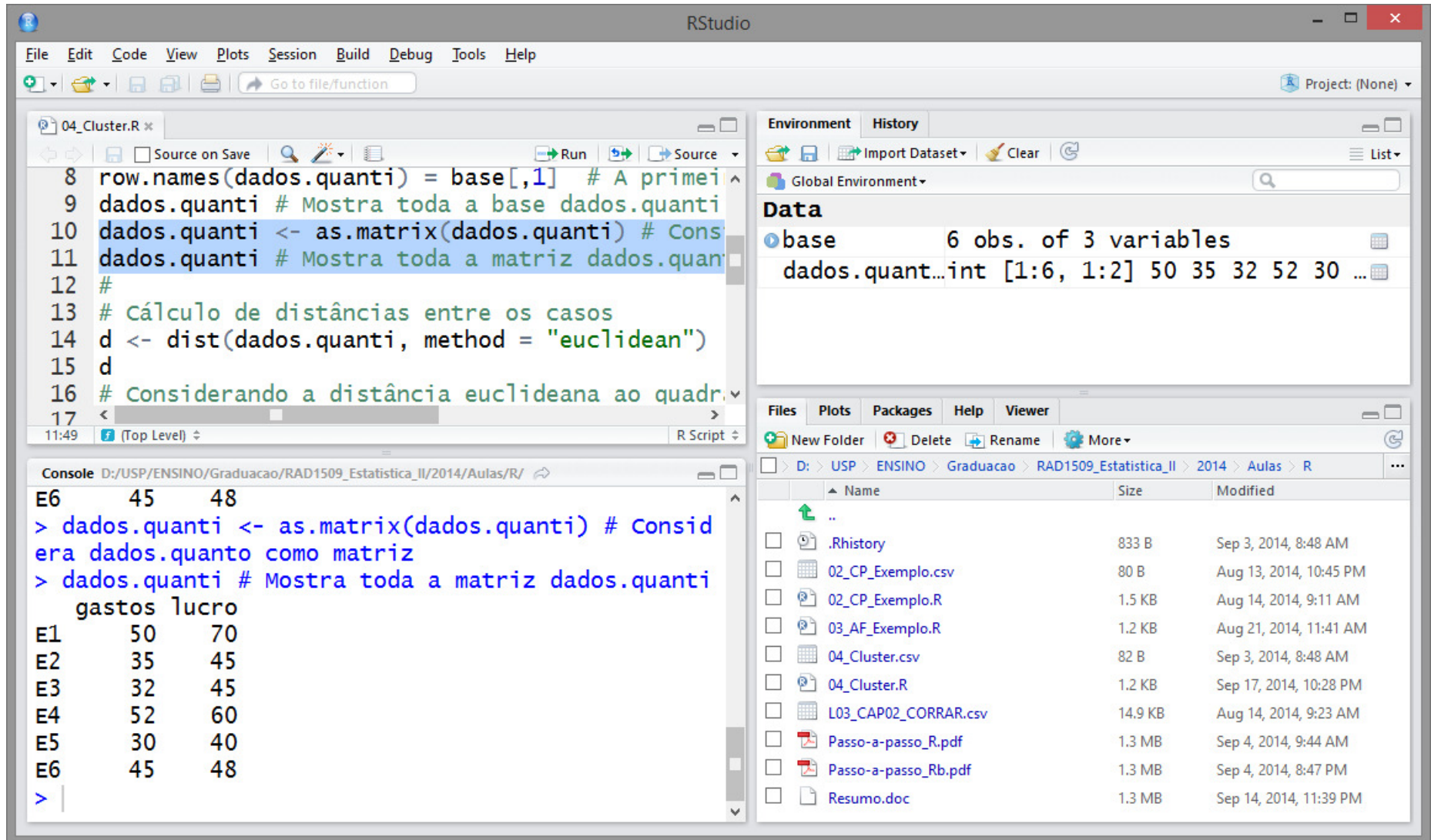
4 base # Mostra toda a base de dados
5 #
6 dados.quanti <- base[,-1] # Remove a primeira
7 dados.quanti # Mostra toda a base dados.quanti
8 row.names(dados.quanti) = base[,1] # A primeira
9 dados.quanti # Mostra toda a base dados.quanti
10 dados.quanti <- as.matrix(dados.quanti) # Converte
11 dados.quanti # Mostra toda a matriz dados.quanti
12 #
13 <
      
```
- Environment:** Shows the Global Environment with two data objects:
 - base:** 6 obs. of 3 variables
 - dados.quanti:** 6 obs. of 2 variables
- Console:** Shows the execution of the code, resulting in a matrix of 6 rows and 2 columns. The first column contains case identifiers (E1-E6).


```

6      45      48
> row.names(dados.quanti) = base[,1] # A primeira
coluna da base contém os casos
> dados.quanti # Mostra toda a base dados.quanti
      gastos lucro
E1      50      70
E2      35      45
E3      32      45
E4      52      60
E5      30      40
E6      45      48
      
```
- Files:** Shows a file explorer view of the project directory, listing files such as .Rhistory, 02_CP_Exemplo.csv, 02_CP_Exemplo.R, 03_AF_Exemplo.R, 04_Cluster.csv, 04_Cluster.R, L03_CAP02_CORRAR.csv, Passo-a-passo_R.pdf, Passo-a-passo_Rb.pdf, and Resumo.doc.

Análise de Agrupamentos no R

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



The screenshot shows the RStudio interface with the following components:

- Source Editor:** Contains R code for converting the data frame to a matrix and calculating distances.


```

8 row.names(dados.quant) = base[,1] # A primeira linha da base é a lista de nomes
9 dados.quant # Mostra toda a base dados.quant
10 dados.quant <- as.matrix(dados.quant) # Converte a base em matriz
11 dados.quant # Mostra toda a matriz dados.quant
12 #
13 # Cálculo de distâncias entre os casos
14 d <- dist(dados.quant, method = "euclidean")
15 d
16 # Considerando a distância euclidiana ao quadrado
17 <
      
```
- Environment:** Shows the 'Data' tab with the 'base' object (6 obs. of 3 variables) and the 'dados.quant' object (int [1:6, 1:2] 50 35 32 52 30 ...).
- Console:** Shows the execution of the code, resulting in the distance matrix 'd'.

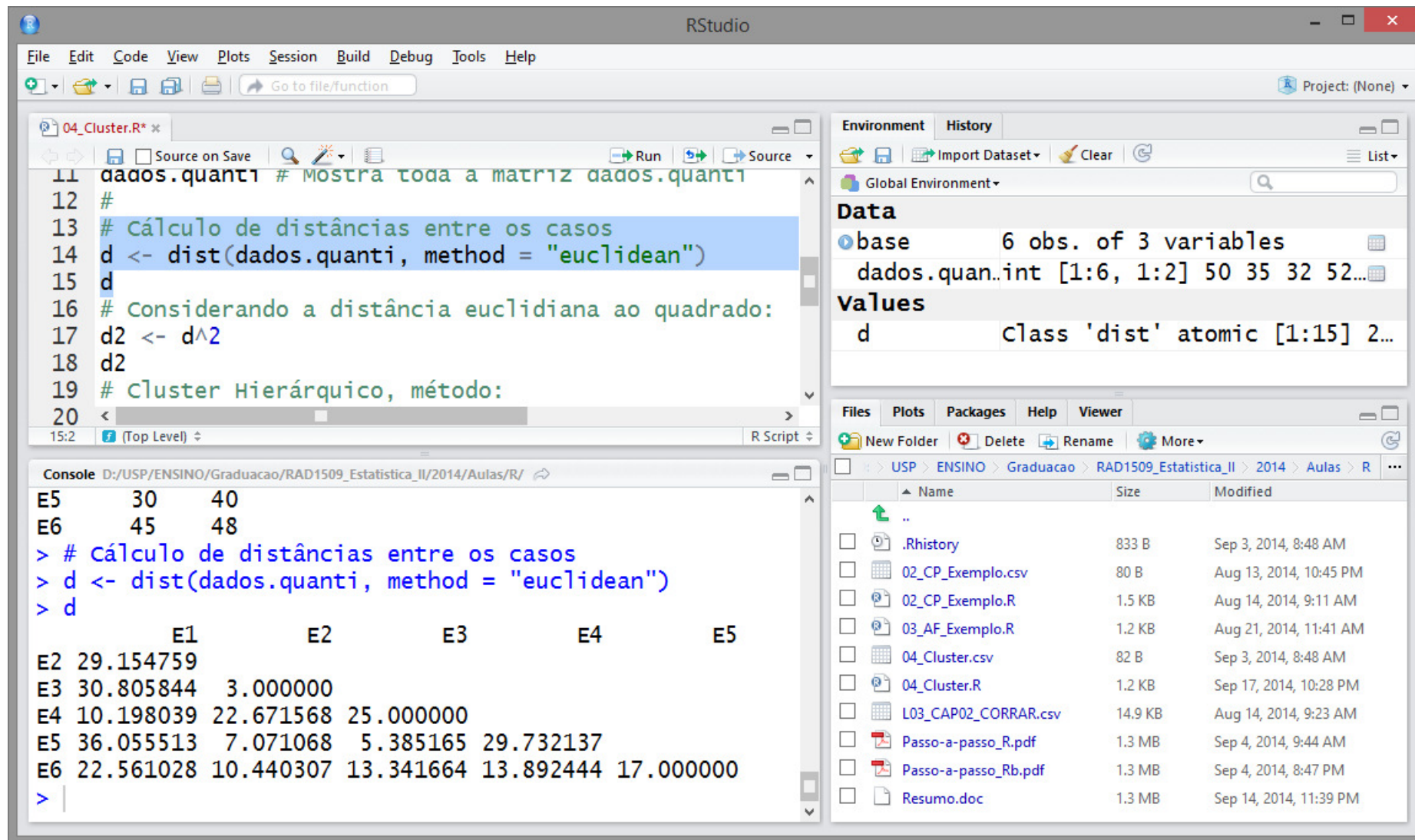

```

E6      45      48
> dados.quant <- as.matrix(dados.quant) # Considera dados.quant como matriz
> dados.quant # Mostra toda a matriz dados.quant
      gastos lucro
E1      50      70
E2      35      45
E3      32      45
E4      52      60
E5      30      40
E6      45      48
      
```
- Files:** Shows the file explorer with the current directory being 'D:/USP/ENSINO/Graduacao/RAD1509_Estatistica_II/2014/Aulas/R'. The file list includes:

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:

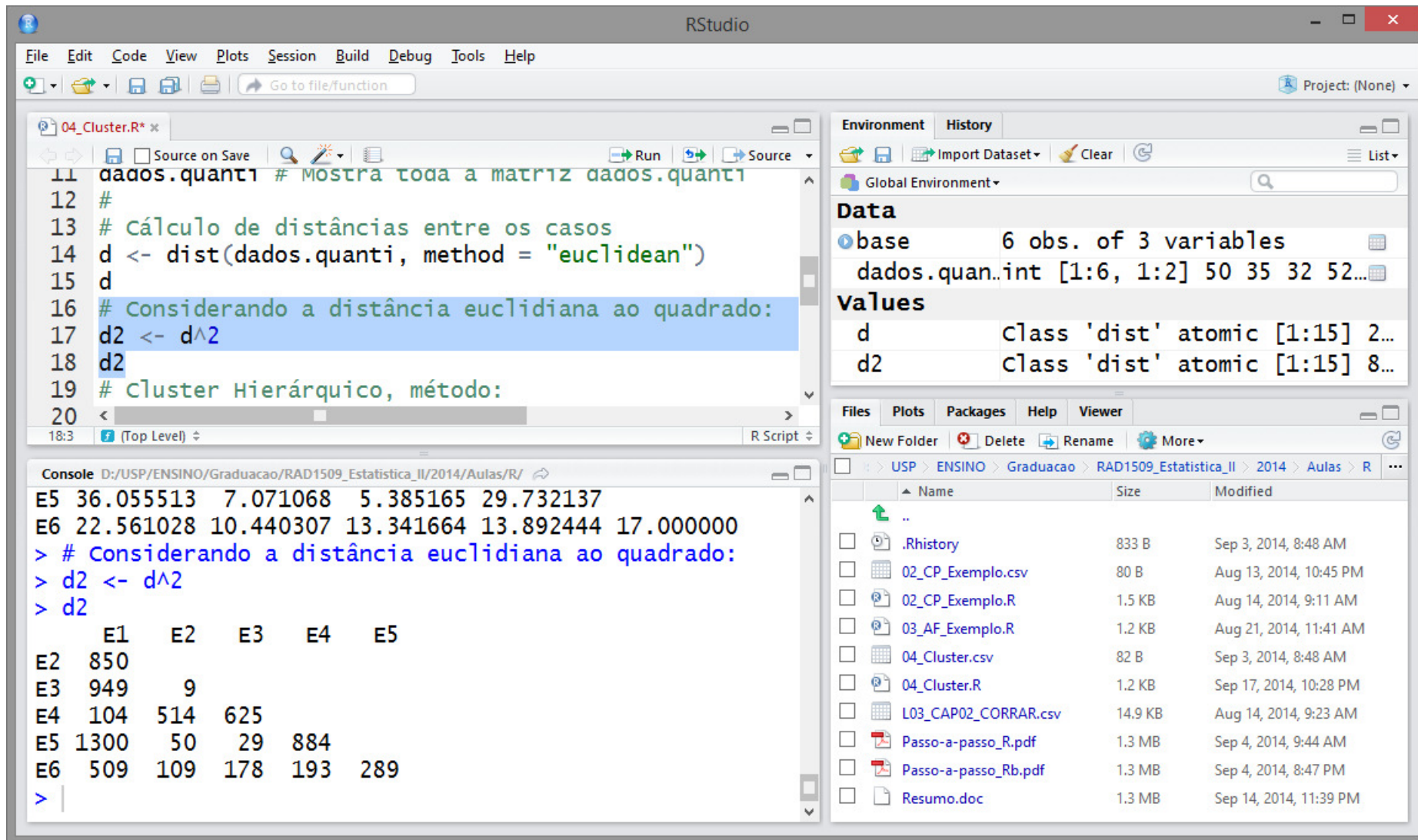
- Source Editor:** Contains R code for calculating a distance matrix. The code is as follows:


```
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 <
```
- Environment:** Shows the global environment with the following data:
 - base:** 6 obs. of 3 variables
 - dados.quan.int [1:6, 1:2]:** 50 35 32 52...
 - values:** d Class 'dist' atomic [1:15] 2...
- Files:** Shows a list of files in the project directory, including .Rhistory, 02_CP_Exemplo.csv, 02_CP_Exemplo.R, 03_AF_Exemplo.R, 04_Cluster.csv, 04_Cluster.R, L03_CAP02_CORRAR.csv, Passo-a-passo_R.pdf, Passo-a-passo_Rb.pdf, and Resumo.doc.
- Console:** Shows the output of the R code, including the calculation of the distance matrix and the resulting matrix of distances between cases E1, E2, E3, E4, and E5.


```
> # 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

Calcule a distância euclidiana ao quadrado.



The screenshot shows the RStudio interface with the following components:

- Source Editor:** Contains R code for calculating Euclidean distance and performing hierarchical clustering. The line `d2 <- d^2` is highlighted in blue.
- Environment:** Shows the global environment with variables `base` (6 obs. of 3 variables) and `dados.quan.int` [1:6, 1:2] 50 35 32 52....
- Files:** Shows a file explorer with various files including `.Rhistory`, `02_CP_Exemplo.csv`, `02_CP_Exemplo.R`, `03_AF_Exemplo.R`, `04_Cluster.csv`, `04_Cluster.R`, `L03_CAP02_CORRAR.csv`, `Passo-a-passo_R.pdf`, `Passo-a-passo_Rb.pdf`, and `Resumo.doc`.
- Console:** Displays the output of the R code, showing the distance matrix `d2` and the hierarchical clustering result.

R Code:

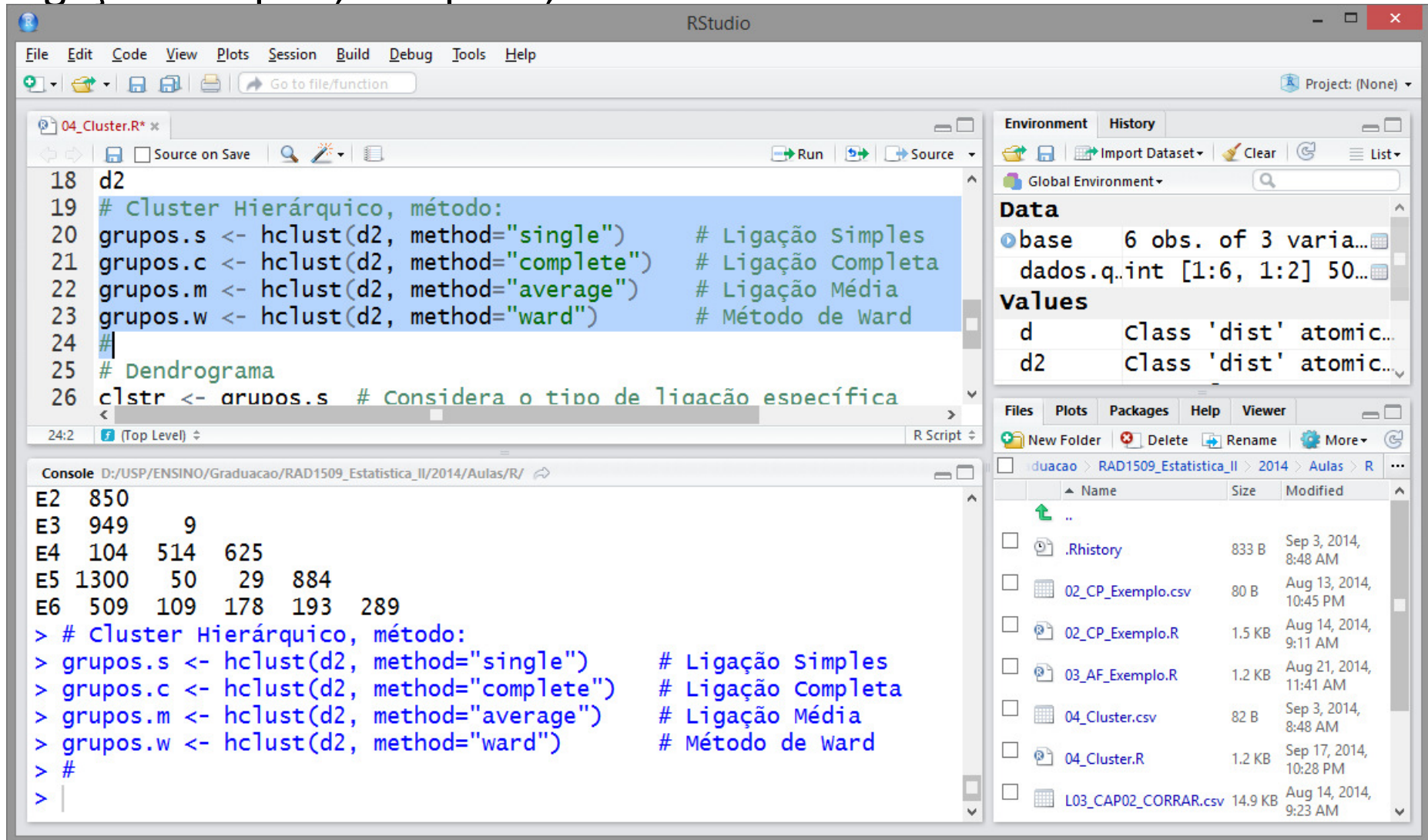
```
11 dados.quanti # Mostra toda a matriz dados.quanti
12 #
13 # Cálculo de distâncias entre os casos
14 d <- dist(dados.quanti, method = "euclidean")
15 d
16 # Considerando a distância euclidiana ao quadrado:
17 d2 <- d^2
18 d2
19 # Cluster Hierárquico, método:
20 <
```

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
>
```

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 a script editor, environment pane, and console.

Script Editor (04_Cluster.R*):

```

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 <- grupos.s # Considera o tipo de ligação específica
  
```

Environment Pane:

- Data:**
 - base: 6 obs. of 3 variables
 - dados.q.int [1:6, 1:2] 50...
- Values:**
 - d: Class 'dist' atomic...
 - d2: Class 'dist' atomic...

Files Pane:

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

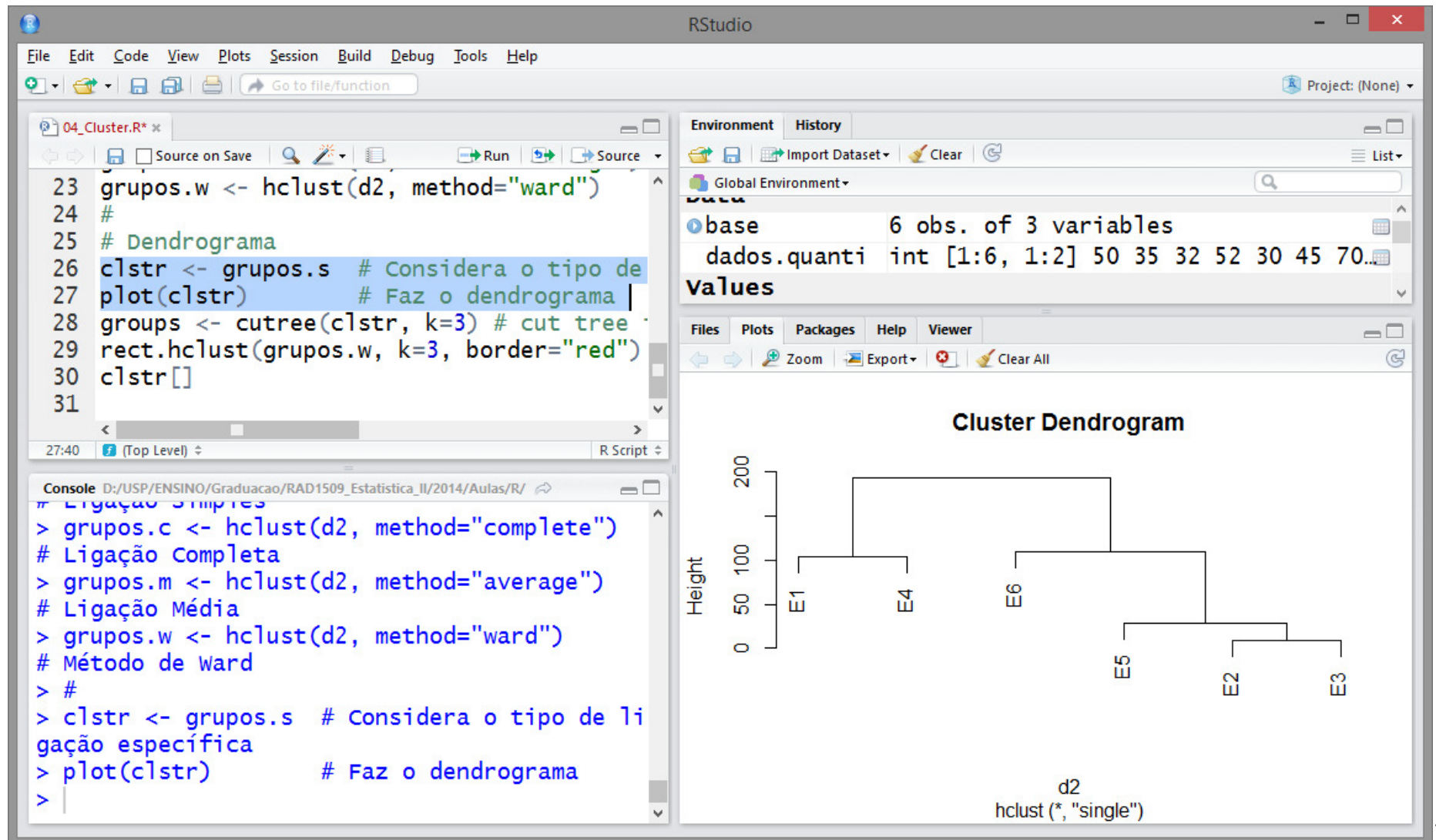
Console:

```

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
> #
>
  
```

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.

RStudio interface showing the execution of R code for hierarchical clustering and the resulting dendrogram.

Code Editor (04_Cluster.R*):

```

22 grupos.m <- hclust(d2, method="average")
23 grupos.w <- hclust(d2, method="ward")
24 #
25 # Dendrograma
26 clstr <- grupos.s # Considera o tipo de
27 plot(clstr)       # Faz o dendrograma
28 groups <- cutree(clstr, k=3) # cut tree
29 rect.hclust(grupos.w, k=3, border="red")
30 clstr[]
31

```

Environment:

Object	Class	Value
groups	Named int [1:6]	1 2 2 1 2 3
grupos.c	List of 7	
grupos.m	List of 7	
grupos.s	List of 7	

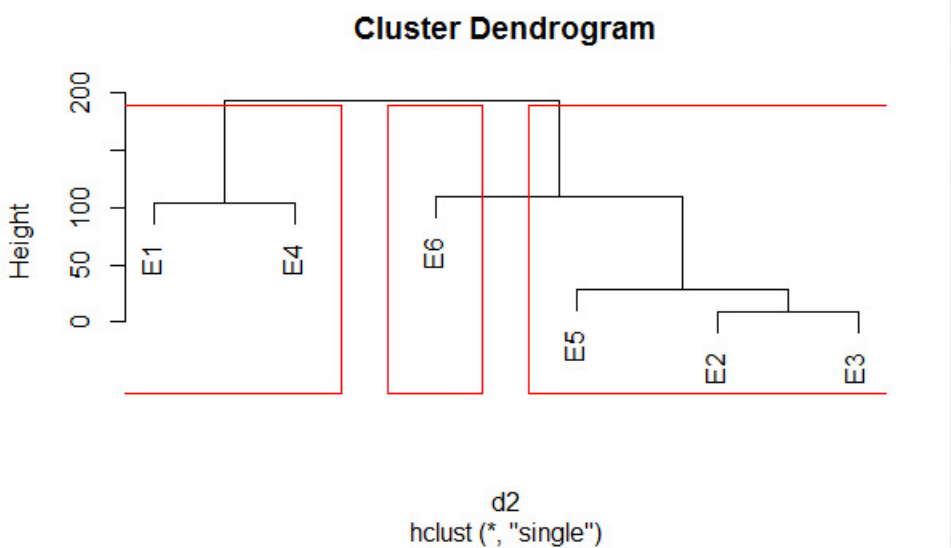
Console:

```

> grupos.w <- hclust(d2, method="ward")
# Método de Ward
> clstr <- grupos.s # Considera o tipo de ligação específica
> plot(clstr)       # Faz o dendrograma
> groups <- cutree(clstr, k=3) # cut tree into 3 clusters
> rect.hclust(grupos.w, k=3, border="red") # insere retangulos no dendrograma
>

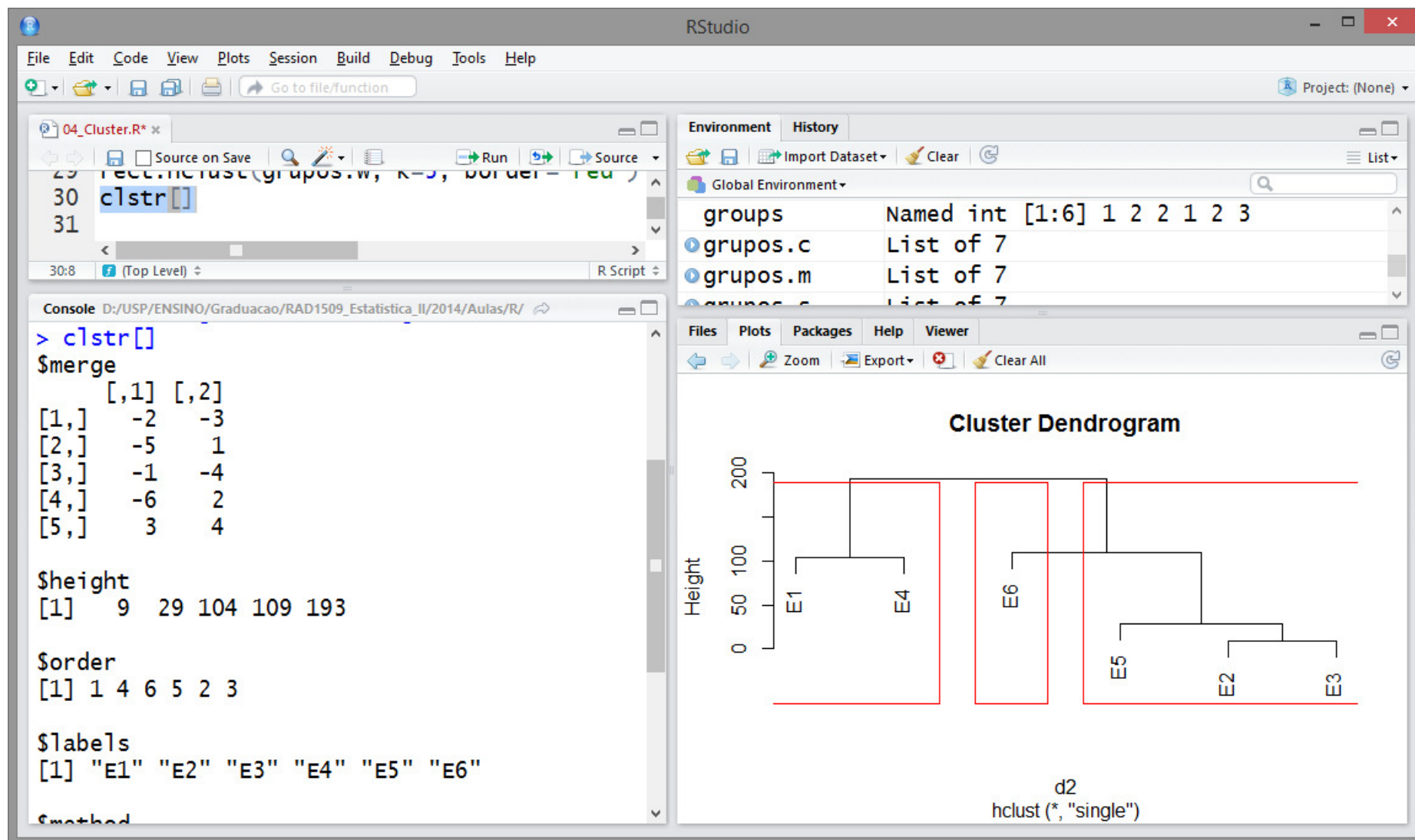
```

Cluster Dendrogram:

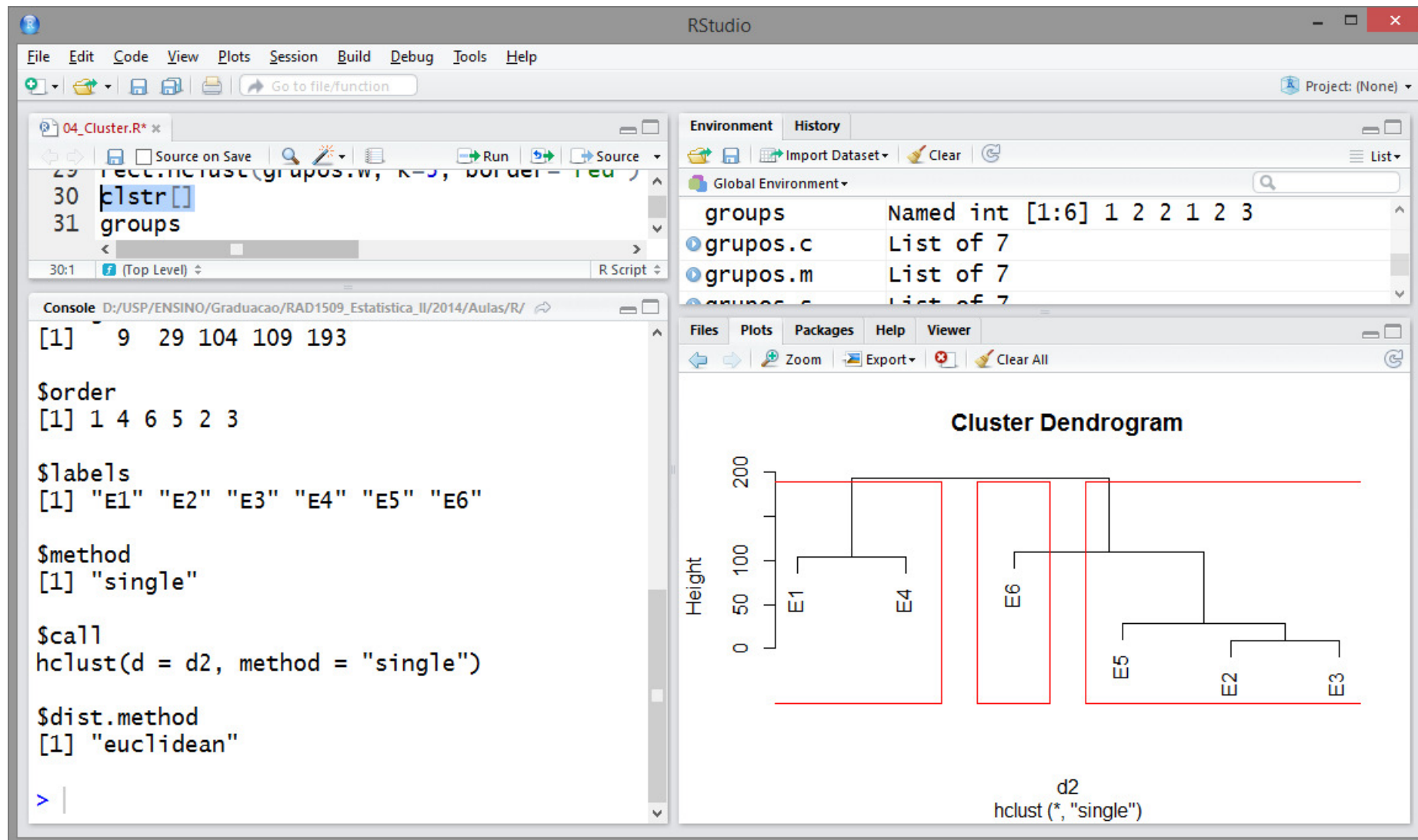


The dendrogram illustrates the hierarchical clustering of 6 samples (E1, E2, E3, E4, E5, E6) based on the Ward method. The y-axis represents the height of the clusters, ranging from 0 to 200. The x-axis is labeled 'd2 hclust (*, "single")'. Red rectangles highlight the three clusters identified by the cutree function (k=3): {E1, E4}, {E6}, and {E2, E3, E5}.

Verifique a análise completa.



Verifique a análise completa.



Análise de Agrupamentos no R

Obtenha a lista dos membros de cada agrupamento

RStudio

File Edit Code View Plots Session Build Debug Tools Help

Go to file/function

Project: (None)

04_Cluster.R*

```

17 d2 <- d^2
18 d2
19 # Cluster Hierárquico, método:
20 grupos.s <- hclust(d2, method="single")
21 grupos.c <- hclust(d2, method="complete")
22 grupos.m <- hclust(d2, method="average")
23 grupos.w <- hclust(d2, method="ward")
24 #
25 # Dendrograma
26 clstr <- grupos.s # Considera o tipo de
27 plot(clstr) # Faz o dendrograma
28 grupos <- cutree(clstr, k=3) # cut tree
29 rect.hclust(grupos.w, k=3, border="red")
30 clstr[]
31 grupos

```

Environment History

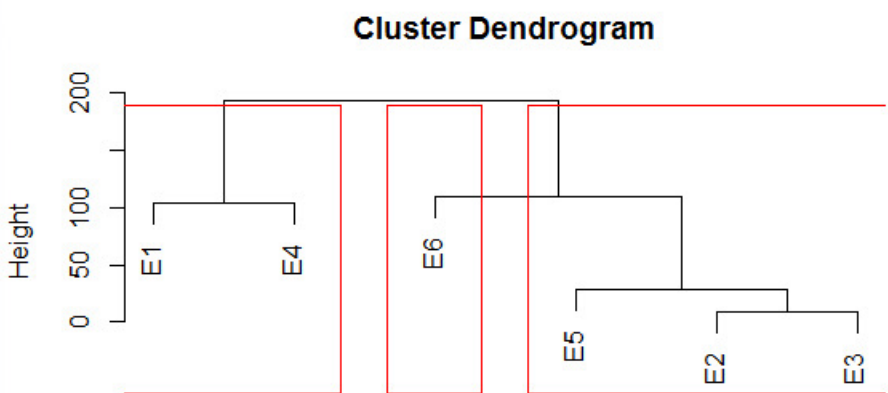
Global Environment

Object	Class	Attributes
grupos	Named int	[1:6] 1 2 2 1 2 3
grupos.c	List of 7	
grupos.m	List of 7	
grupos.s	List of 7	
grupos.w	List of 7	

Files Plots Packages Help Viewer

Zoom Export Clear All

Cluster Dendrogram



Height

d2
hclust (*, "single")

Console

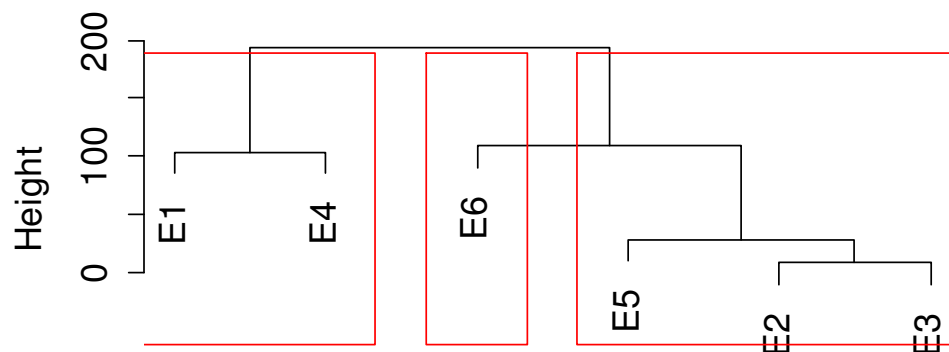
```

> grupos
E1 E2 E3 E4 E5 E6
1 2 2 1 2 3
>

```

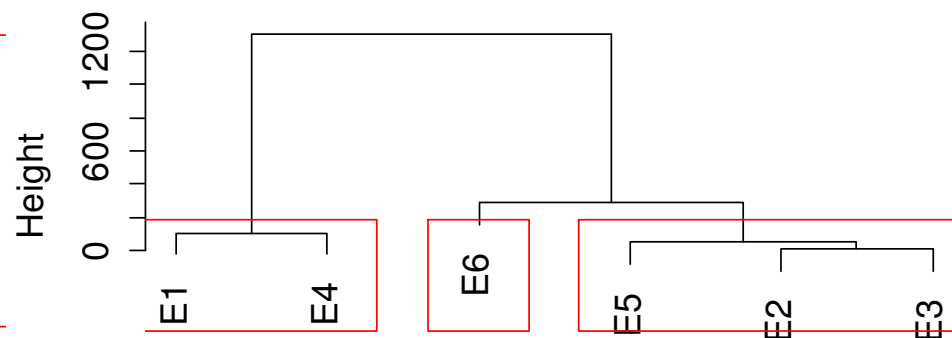

Análise de Agrupamentos no R

Cluster Dendrogram



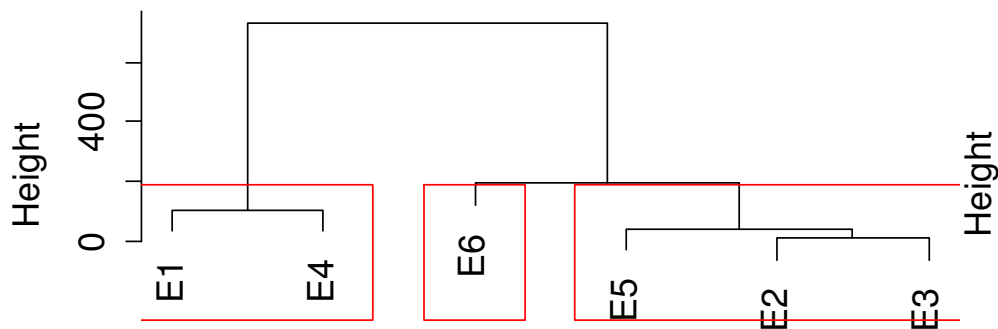
`d2`
`hclust (*, "single")`

Cluster Dendrogram



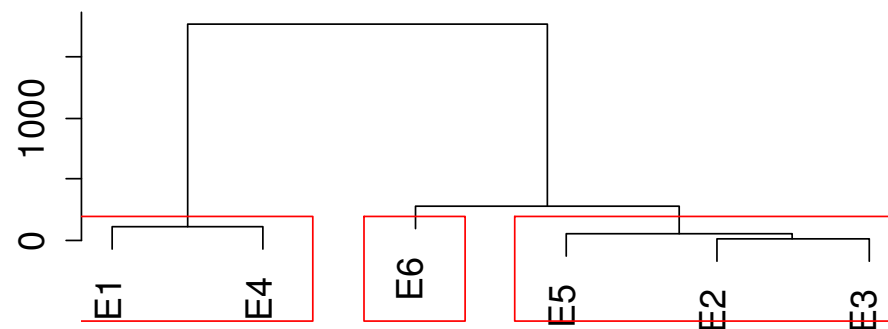
`d2`
`hclust (*, "complete")`

Cluster Dendrogram



`d2`
`hclust (*, "average")`

Cluster Dendrogram



`d2`
`hclust (*, "ward")`