

UNIVERSIDAD LAICA ELOY ALFARO DE MANABI
FACULTAD DE CIENCIAS INFORMATICAS
CARRERA DE TECNOLOGIA DE LA INFORMACION

TEMA:

ANÁLISIS DE CONGLOMERADOS

NOMBRE:

MACIAS PICO JOSSELYN STEFANY

CURSO:

SEXTO “B”

MATERIA:

MINERIA DE DATOS

DOCENTE:

ING. FABRICIO RIVADENEIRA

FECHA:

26-07-2021

MANTA-MANABI-ECUADOR

Introducción

El análisis de conglomerados es una técnica multivariante que permite agrupar los casos o variables de un archivo de datos en función del parecido o similaridad existente entre ellos. El objetivo fundamental de un análisis de conglomerados es realizar una partición de la muestra en grupos similares el punto de partida es una matriz de similaridad o de distancias entre los sujetos, objetos o variables que se desean agrupar. Las etapas en un análisis de conglomerados utilizando primero una matriz de datos en el cual se obtiene los segmentos, después una matriz de similaridad en esta se realiza un perfilado de segmentos, por último, un algoritmo de clasificación.

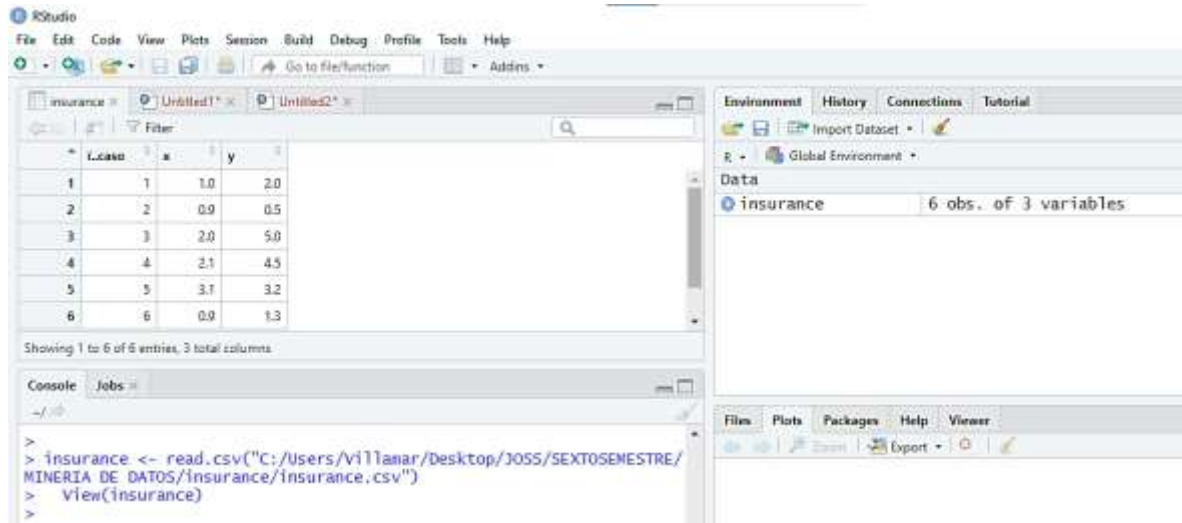
Como una técnica de agrupación de variables, el análisis de conglomerados es similar al análisis factorial. La aglomeración es menos restrictiva en sus supuestos y admite varios métodos de estimación de la matriz de distancias. Los conglomerados tienen un análisis estadístico el cual necesita una medida para calcular la distancia entre dos sujetos, lo distintos que son y un criterio, una regla un método para agruparlos y asignarlos a cada conglomerado.

Mientras que el análisis discriminante efectúa la clasificación tomando como referencia un criterio o variable dependiente, el análisis de conglomerados permite detectar el numero optimo de grupos y su composición únicamente a partir de la similaridad existente entre los casos. Como técnica de agrupación de casos, el análisis de conglomerados es similar y no asume ninguna distribución específica para las variables.

Desarrollo de instrucciones

1. Considerar el conjunto de puntos siguientes: (1.0; 2.0) (0.9; 0.5) (2.0; 5.0) (2.1; 4.5) (3.1; 3.2) (0.9; 1.3)

- Importar los datos de los puntos en x, y para utilizarlos a medida que el proceso continuo

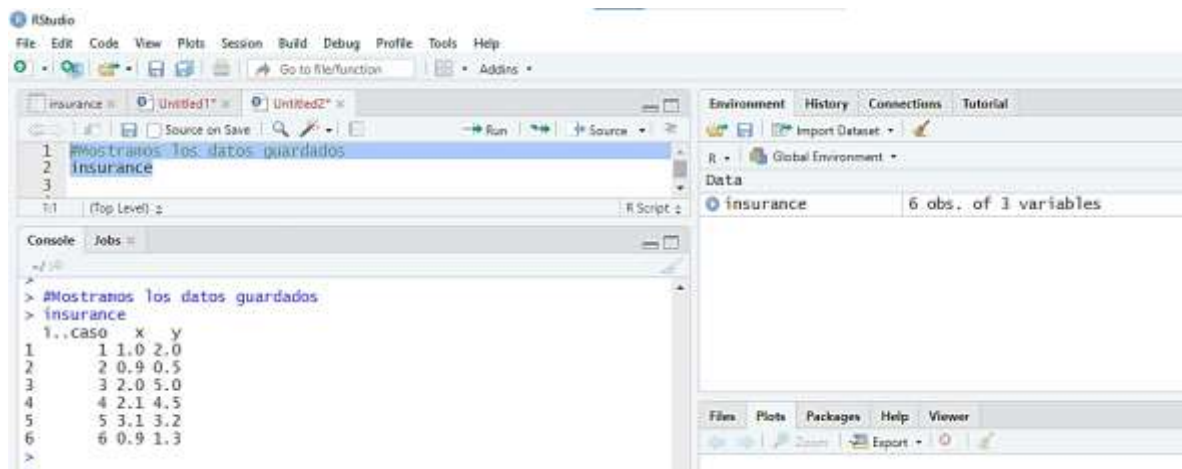


```
> insurance <- read.csv("C:/Users/villanar/Desktop/JOSS/SEXTOSEMESTRE/MINERIA DE DATOS/insurance/insurance.csv")
> View(insurance)
```

The screenshot shows the RStudio interface. The Environment pane on the right displays the 'insurance' dataset with 6 observations and 3 variables. The Data viewer shows the first few rows of the dataset.

	1	2	3	4	5	6
1	1	1.0	2.0			
2	2	0.9	0.5			
3	3	2.0	5.0			
4	4	2.1	4.5			
5	5	3.1	3.2			
6	6	0.9	1.3			

- Se muestran los datos guardados para verificar antes de usarlos

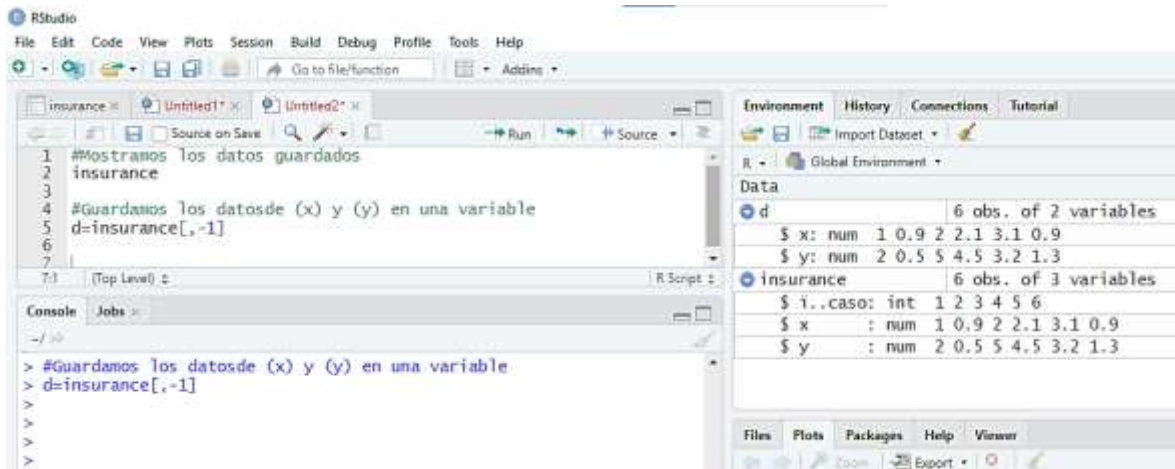


```
> #Mostramos los datos guardados
> insurance
```

The screenshot shows the RStudio interface. The Console pane displays the command to view the 'insurance' dataset. The Environment pane on the right shows the 'insurance' dataset with 6 observations and 3 variables. The Data viewer shows the first few rows of the dataset.

	1	2	3	4	5	6
1	1	1.0	2.0			
2	2	0.9	0.5			
3	3	2.0	5.0			
4	4	2.1	4.5			
5	5	3.1	3.2			
6	6	0.9	1.3			

- Guardamos los datos de (x) y (y) en una variable la cual esta denominada **d**



```

1 #Mostramos los datos guardados
2 insurance
3
4 #Guardamos los datos de (x) y (y) en una variable
5 d=insurance[,-1]
6
7

```

Console:

```

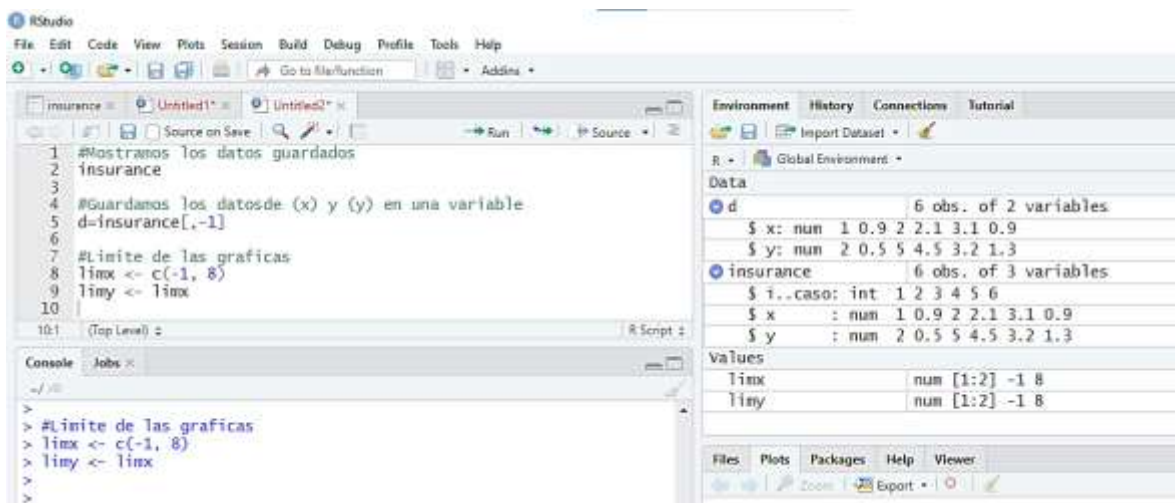
> #Guardamos los datos de (x) y (y) en una variable
> d=insurance[,-1]
>
>
>

```

Environment:

Object	Class	Attributes
d	data.frame	6 obs. of 2 variables
\$ x	num	1 0.9 2 2.1 3.1 0.9
\$ y	num	2 0.5 5 4.5 3.2 1.3
insurance	data.frame	6 obs. of 3 variables
\$ i.caso	int	1 2 3 4 5 6
\$ x	num	1 0.9 2 2.1 3.1 0.9
\$ y	num	2 0.5 5 4.5 3.2 1.3

- Luego se le ubica un limite a las graficas tanto de x como de y



```

1 #Mostramos los datos guardados
2 insurance
3
4 #Guardamos los datos de (x) y (y) en una variable
5 d=insurance[,-1]
6
7 #Limite de las graficas
8 limx <- c(-1, 8)
9 limy <- limx
10

```

Console:

```

> #limite de las graficas
> limx <- c(-1, 8)
> limy <- limx
>
>

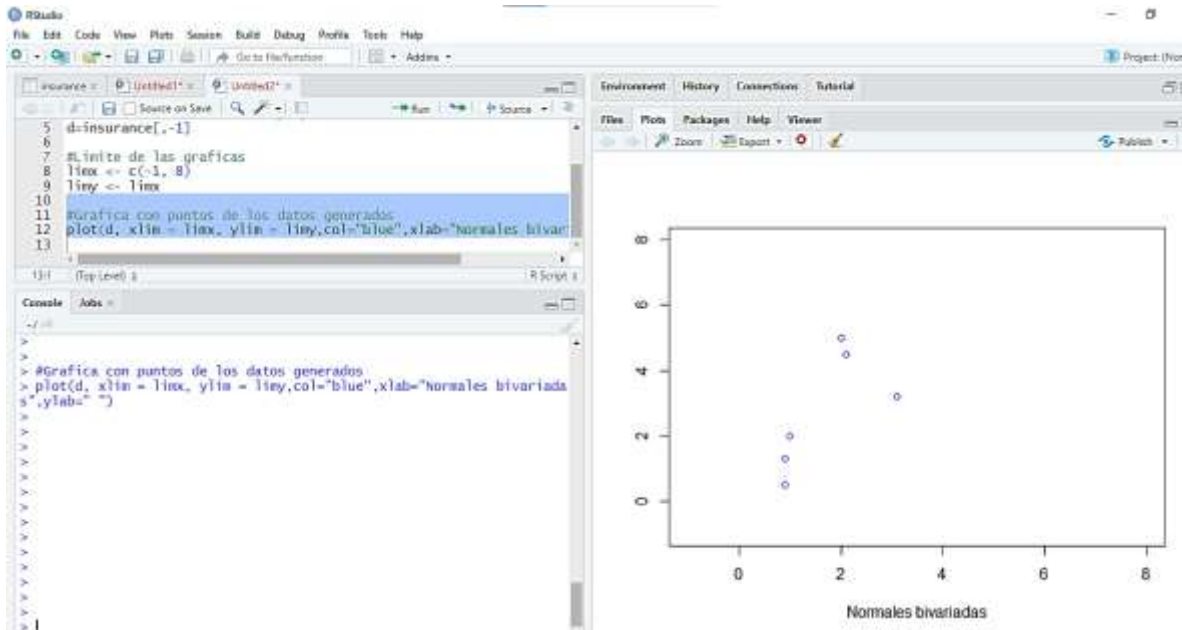
```

Environment:

Object	Class	Attributes
d	data.frame	6 obs. of 2 variables
\$ x	num	1 0.9 2 2.1 3.1 0.9
\$ y	num	2 0.5 5 4.5 3.2 1.3
insurance	data.frame	6 obs. of 3 variables
\$ i.caso	int	1 2 3 4 5 6
\$ x	num	1 0.9 2 2.1 3.1 0.9
\$ y	num	2 0.5 5 4.5 3.2 1.3
limx	num	[1:2] -1 8
limy	num	[1:2] -1 8

2. Presentar una gráfica de los puntos en el plano x-y (gráfico de dispersión).

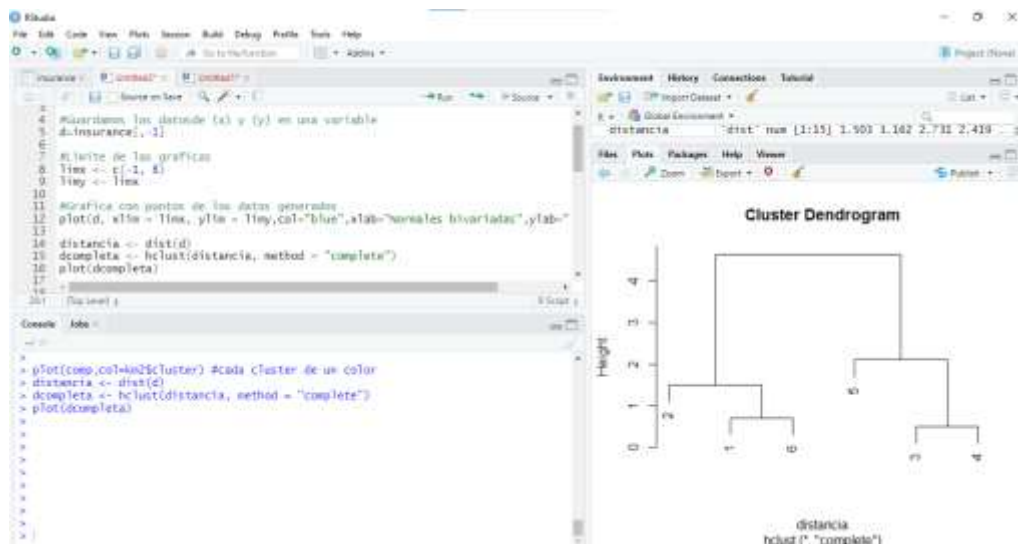
- Después se muestra los datos los puntos dentro de una grafica



3. Presentar al menos 2 diferentes agrupamientos jerárquicos del dataset. (2 diferentes dendrogramas) y comentar sobre alguna comparación o diferencia.

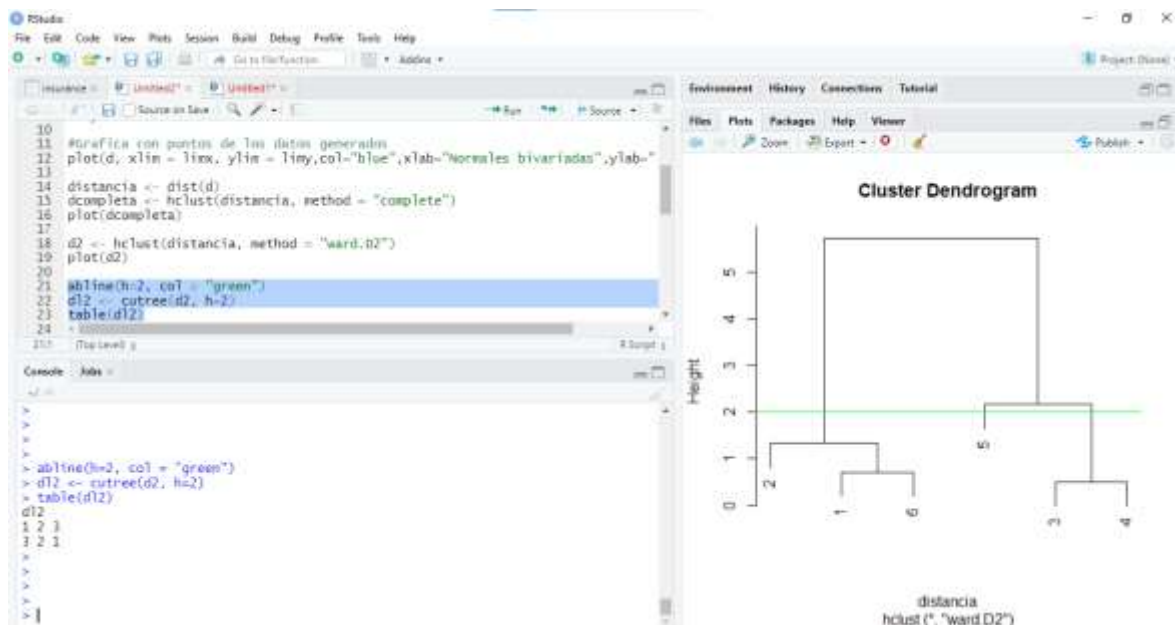
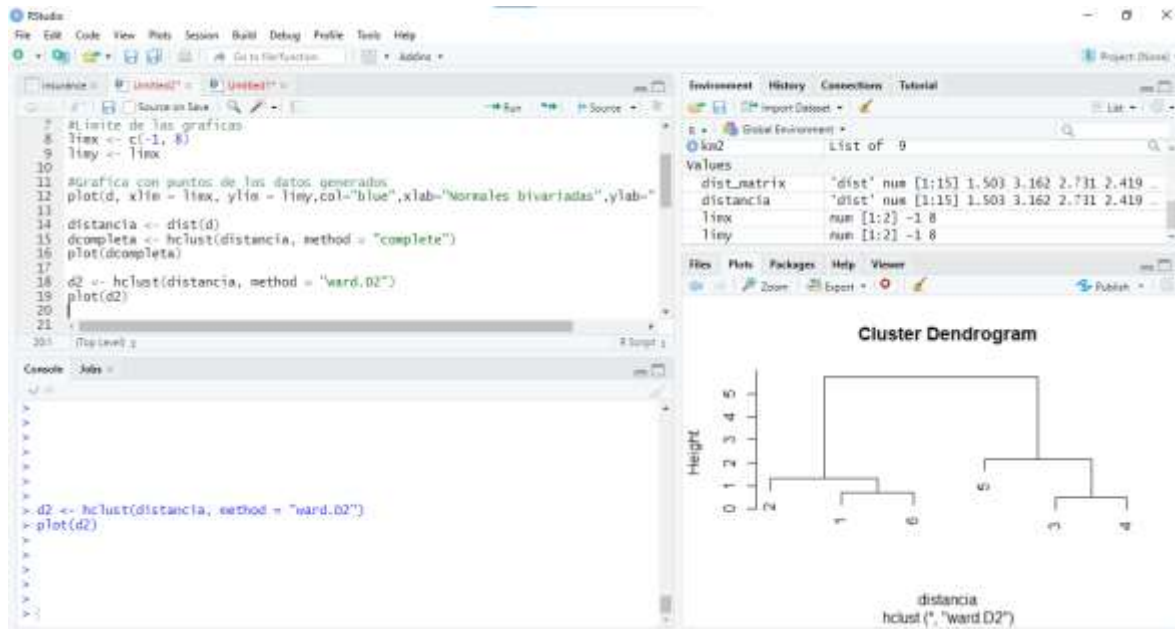
- **Método Complete:**

Este método también se llama diámetro o método máximo. En este método, consideramos la similitud del par más lejano. Es decir, la distancia entre un clúster y otro clúster se considera igual a la distancia más larga desde cualquier miembro de un clúster a cualquier miembro del otro clúster. Tiende a producir clústeres más compactos. Una desventaja de este método es que los valores atípicos pueden provocar la fusión de grupos cercanos más tarde de lo que es óptimo.



- **Método Ward**

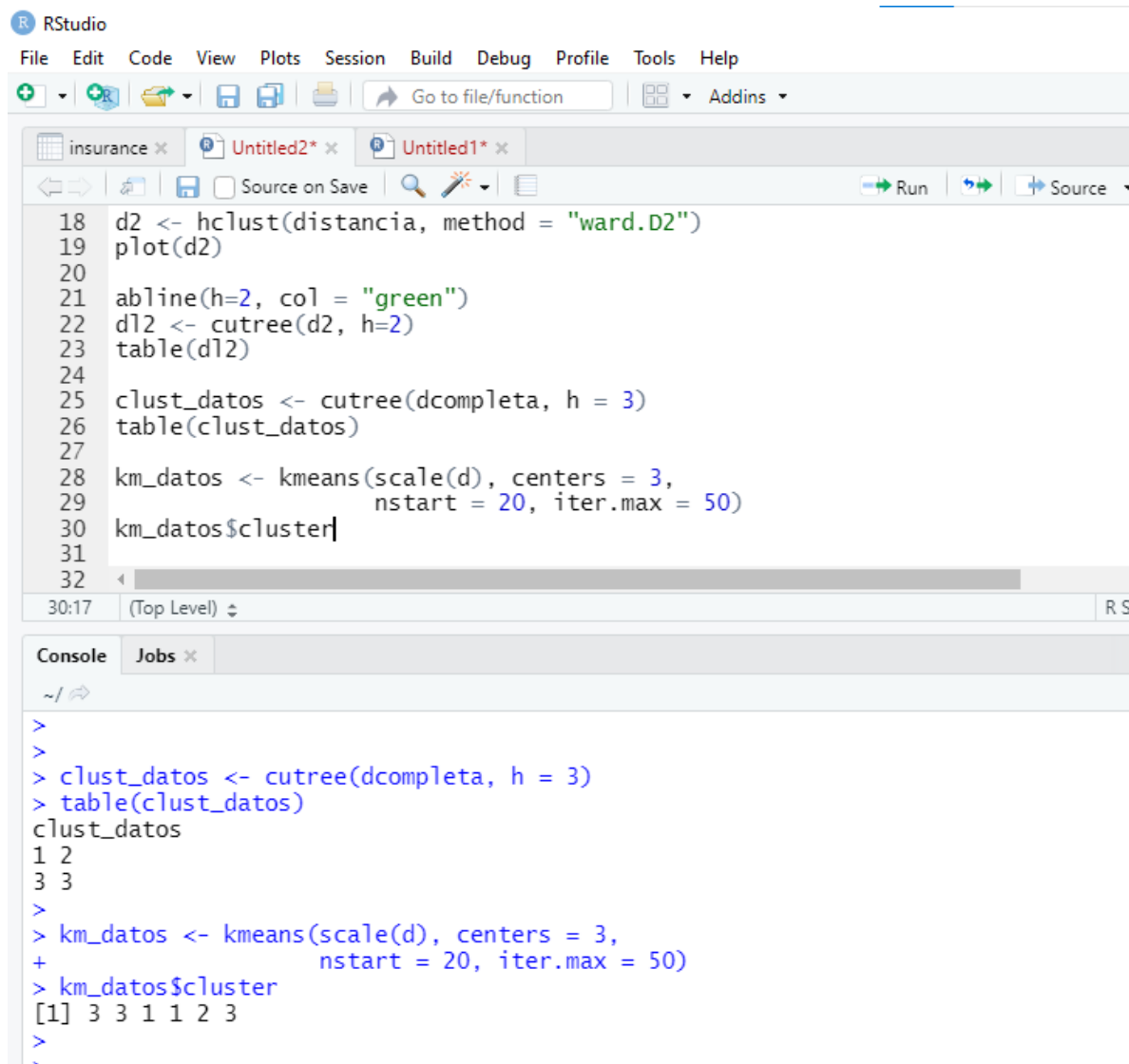
El método de Ward apunta a minimizar la varianza total dentro del grupo. En cada paso, se fusionan el par de clústeres con una distancia mínima entre los clústeres. En otras palabras, forma grupos de una manera que minimiza la pérdida asociada con cada grupo. En cada paso, se considera la unión de cada par de clústeres posible y se combinan los dos clústeres cuya fusión da como resultado un aumento mínimo en la pérdida de información.



- **Comparaciones o diferencias**

Una comparación entre los grupos de métodos ms utilizados son los denominados clúster, cuya importancia reside en la capacidad de agrupar datos, que aparentemente no tienen relación, en la misma clase, por lo que se pueden utilizar en multitud de campos donde se quieren clasificar grandes grupos muestrales.

4. Seleccionar un número de clusters, y ejecutar el agrupamiento por k-means



```
18 d2 <- hclust(distancia, method = "ward.D2")
19 plot(d2)
20
21 abline(h=2, col = "green")
22 d12 <- cutree(d2, h=2)
23 table(d12)
24
25 clust_datos <- cutree(dcompleta, h = 3)
26 table(clust_datos)
27
28 km_datos <- kmeans(scale(d), centers = 3,
29                   nstart = 20, iter.max = 50)
30 km_datos$cluster
31
32
```

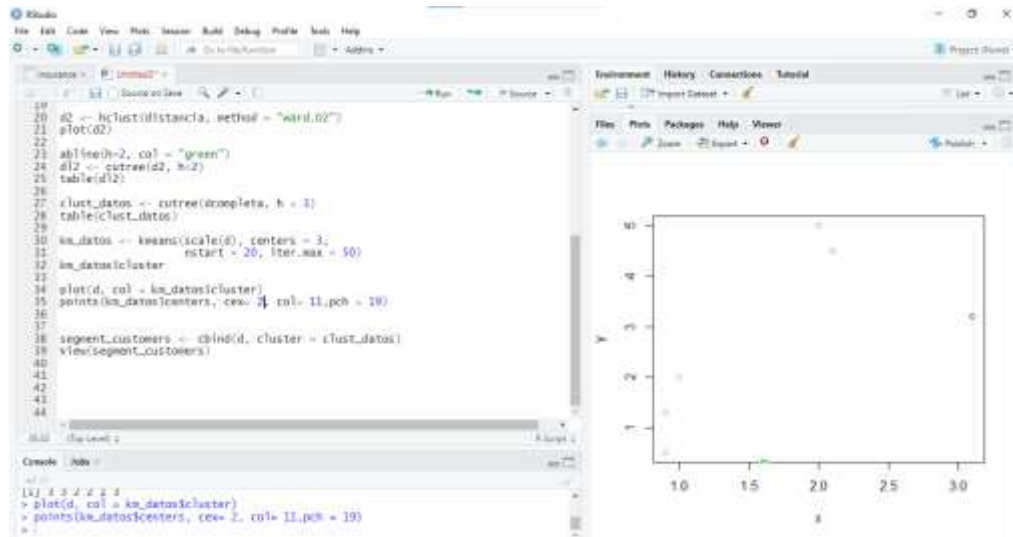
30:17 (Top Level) R S

Console **Jobs**

```
~/
>
>
> clust_datos <- cutree(dcompleta, h = 3)
> table(clust_datos)
clust_datos
1 2
3 3
>
> km_datos <- kmeans(scale(d), centers = 3,
+                   nstart = 20, iter.max = 50)
> km_datos$cluster
[1] 3 3 1 1 2 3
>
>
```

5. Presentar los centroides de cada uno del clúster.

- **Modo Grafico**



- **Otro modo**

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

insurance x Untitled2* x segment_customers x

Filter

	x	y	cluster
1	1.0	2.0	1
2	0.9	0.5	1
3	2.0	5.0	2
4	2.1	4.5	2
5	3.1	3.2	2
6	0.9	1.3	1

Showing 1 to 6 of 6 entries, 3 total columns

Console Jobs x

```

~/
1 2
3 3
> km_datos <- kmeans(scale(d), centers = 3,
+                     nstart = 20, iter.max = 50)
> km_datos$cluster
[1] 3 3 2 2 1 3
> plot(d, col = km_datos$cluster)
> points(km_datos$centers, cex= 2, col= 11,pch = 19)
> segment_customers <- cbind(d, cluster = clust_datos)
> View(segment_customers)
>
>

```


Conclusiones

Después de explicar un procedimiento para formar conglomerados a partir de una serie de datos y también el diagrama en el que se pueden identificar los conglomerados que en un estudio de mercado representa los posibles segmentos a considerar, podemos observar que esta representación gráfica resulta clara para interpretar.

Además de tener aplicación este procedimiento en la segmentación del mercado, lo podríamos aplicar a otras situaciones en las cuales nos interesa identificar semejanzas o diferencias entre preferencias, identificar oportunidades para nuevos productos, identificar nuevos nichos en el mercado, etc. Además del método que se describió, podemos darnos cuenta de que hay otras formas de realizar este análisis y combinando estos y la experiencia podríamos seleccionar el que mejor resuelva el problema de interés