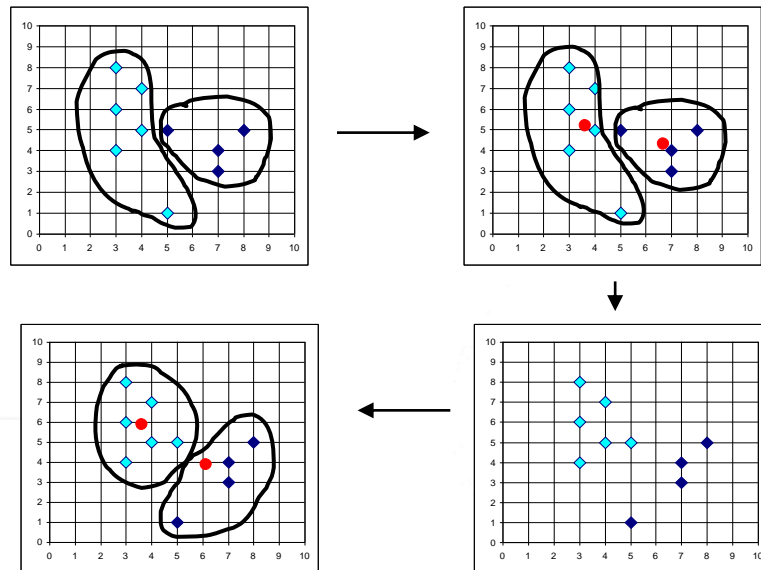




# Método k-medias

[Una introducción]

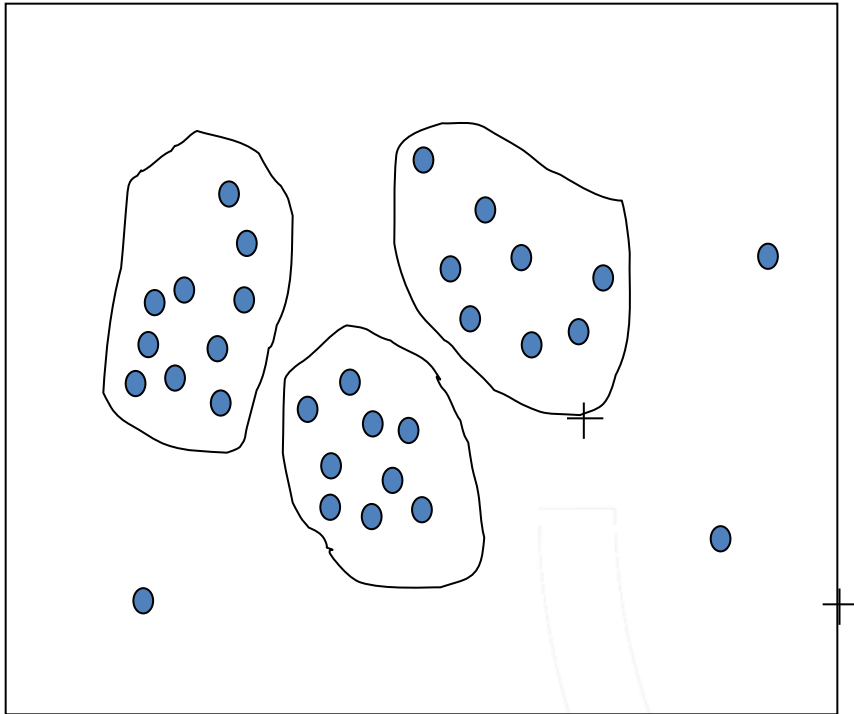
# Método K-Means (Nubes Dinámicas)



# Tareas de la Minería de Datos

- **“Clustering”**: (clasificación no supervisada, aprendizaje no supervizado): Es similar a la clasificación (discriminación), excepto que los grupos no son predefinidos. El objetivo es particionar o segmentar un conjunto de datos o individuos en grupos que pueden ser disjuntos o no. Los grupos se forman basados en la similitud de los datos o individuos en ciertas variables. Como los grupos no son dados a priori el experto debe dar una interpretación de los grupos que se forman.
- **Métodos**:
  - Clasificación Jerárquica (grupos disjuntos).
  - Nubes Dinámicas – k-means (grupos disjuntos).
  - Clasificación Piramidal (grupos NO disjuntos).

# Análisis de Conglomerados

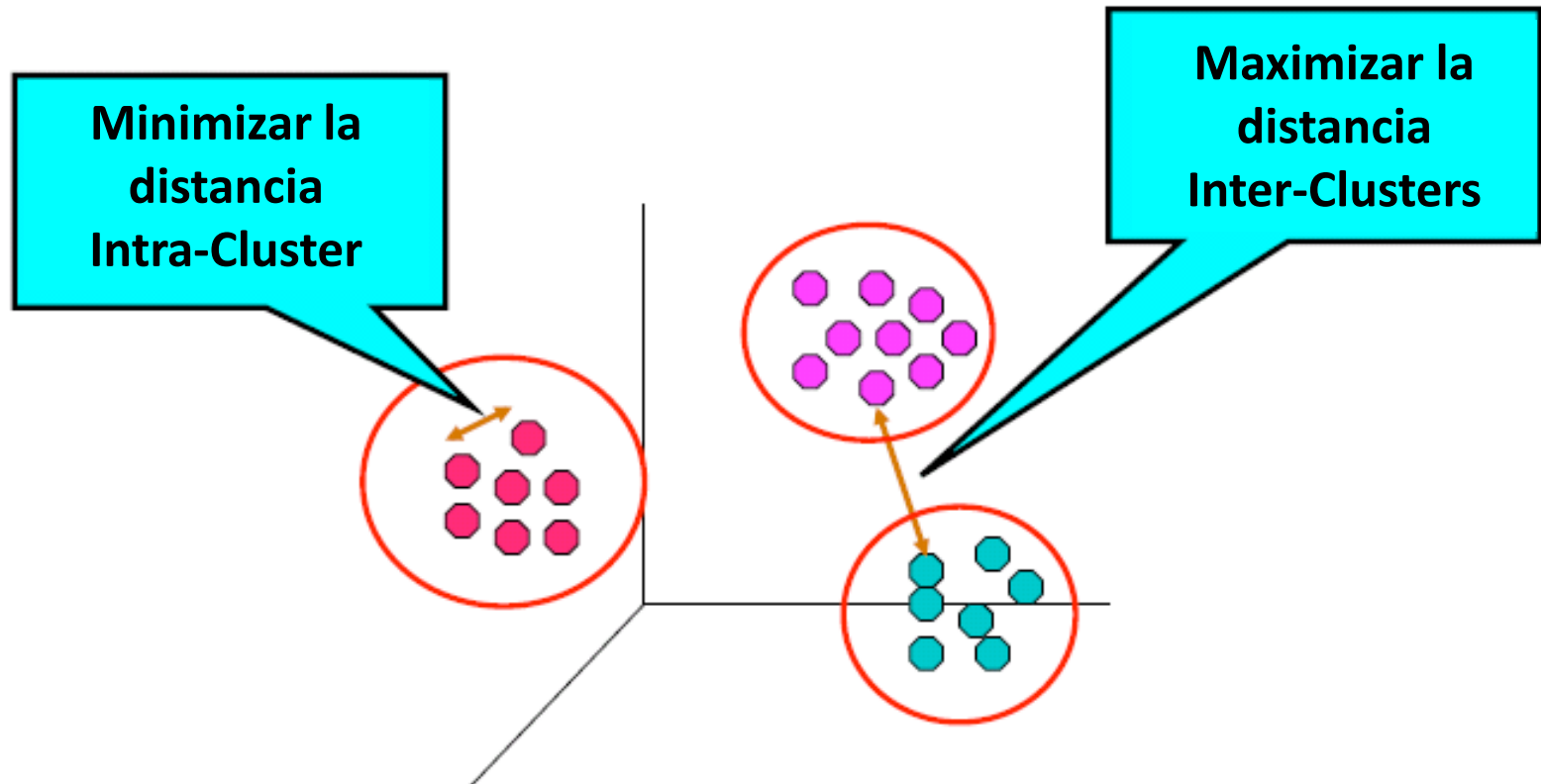


## **Objetivo:**

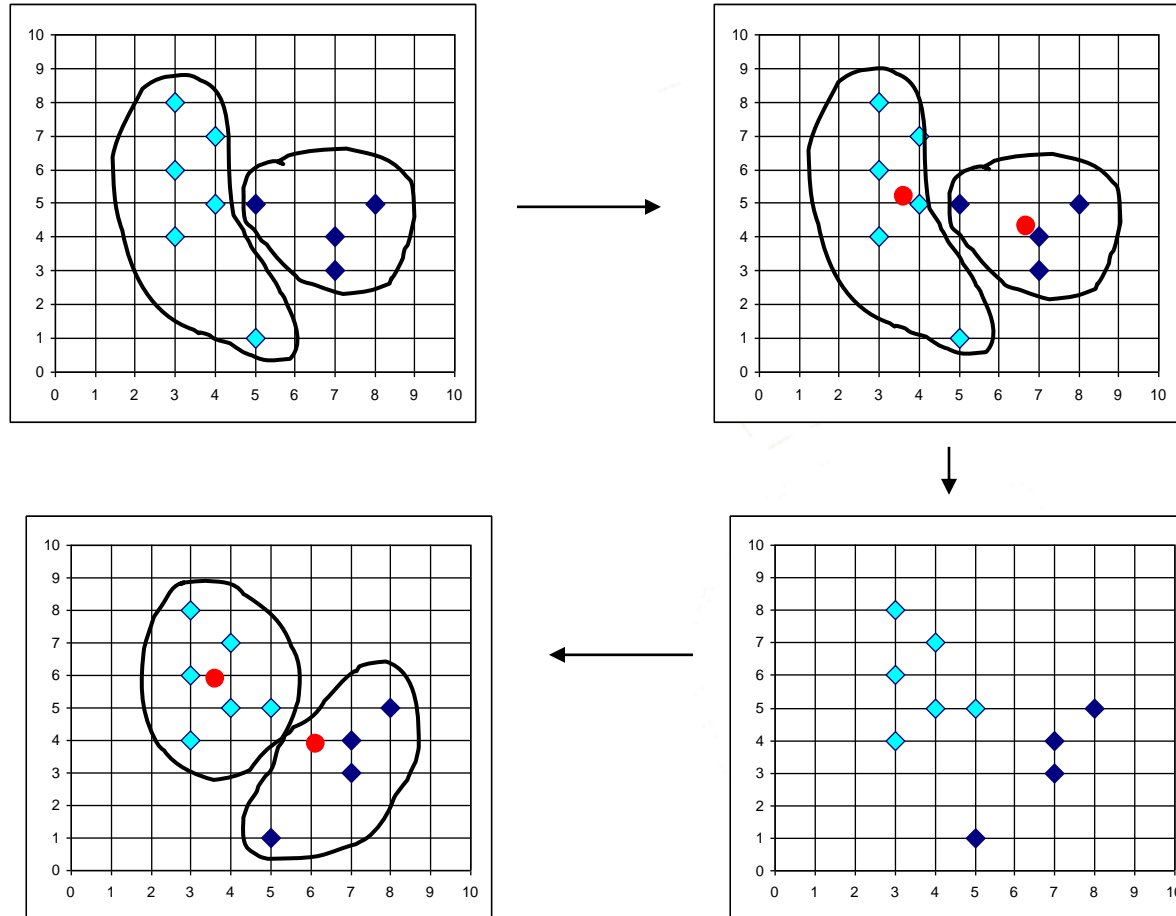
Obtener clases lo más homogéneas posibles y tal que estén suficientemente separadas.



# Criterio de la Inercia



# The *K-Means* Clustering Method (nubes dinámicas)



# Criterio de la inercia

Como se ha mencionado, se quiere obtener clases lo más homogéneas posibles y tal que estén suficientemente separadas. Este objetivo se puede concretar numéricamente a partir de la siguiente propiedad:

supóngase que se está en presencia de una partición  $P = (C_1, C_2, \dots, C_K)$  de  $\Omega$ , donde  $g_1, g_2, \dots, g_K$  son los centros de gravedad de las clases:

$$g_k = \frac{1}{|C_k|} \sum_{i \in C_k} x_i,$$

$g$  es el centro de gravedad total:

$$g = \frac{1}{n} \sum_{i=1}^n x_i$$



# Ejemplo: Estudiantes

Ver

NotasEscolaresExcelKMeans.xlsx

Análisis de los Clústeres					
	Matemáticas	Ciencias	Español	Historia	EdFísica
Lucía	7	6.5	9.2	8.6	8
Pedro	7.5	9.4	7.3	7	7
Inés	7.6	9.2	8	8	7.5
Luis	5	6.5	6.5	7	9
Andrés	6	6	7.8	8.9	7.3
Ana	7.8	9.6	7.7	8	6.5
Carlos	6.3	6.4	8.2	9	7.2
José	7.9	9.7	7.5	8	6
Sonía	6	6	6.5	5.5	8.7
María	6.8	7.2	8.7	9	7
Centro Gravedad Total de la Nube de Puntos					
	Matemáticas	Ciencias	Español	Historia	EdFísica
	6.79	7.65	7.74	7.9	7.42
Centro Gravedad C1={Pedro,Inés,Ana,José}					
	Matemáticas	Ciencias	Español	Historia	EdFísica
	7.7	9.475	7.625	7.75	6.75
Centro Gravedad C2={Luis,Sonía}					
	Matemáticas	Ciencias	Español	Historia	EdFísica
	5.5	6.25	6.5	6.25	8.85
Centro Gravedad C3={Lucía,Andrés,Carlos,María}					
	Matemáticas	Ciencias	Español	Historia	EdFísica
	6.525	6.525	8.475	8.875	7.375





# Definiciones

- *Inercia total* de la nube de puntos:

$$I = \frac{1}{n} \sum_{i=1}^n ||\mathbf{x}_i - \mathbf{g}||^2$$

- *Inercia inter-clases*, es decir la inercia de los centros de gravedad respecto al centro de gravedad total:

$$B(P) = \sum_{k=1}^K \frac{|C_k|}{n} ||\mathbf{g}_k - \mathbf{g}||^2$$



- *Inercia intra-clases*, es decir la inercia al interior de cada clase:

$$W(P) = \sum_{k=1}^K I(C_k) = \frac{1}{n} \sum_{k=1}^K \sum_{i \in C_k} \|\mathbf{x}_i - \mathbf{g}_k\|^2$$

# Teorema: Igualdad de Fisher

- *Inercia total = Inercia inter-clases*  
+  
*Inercia intra-clases*

$$I = B(P) + W(P)$$



Ver NotasEscolaresExcelKMeans.xlsx



- **Objetivo:** Se quiere que  $B(P)$  sea máxima y  $W(P)$  sea mínima
- Como la inercia  $I(P)$  es fija, dada la nube de puntos, entonces al maximizar  $B(P)$  se minimiza automáticamente  $W(P)$ .
- Por lo tanto, los dos objetivos (homogeneidad al interior de las clases y separación entre las clases) se alcanzan al mismo tiempo al querer minimizar  $W(P)$ .

# Problema combinatorio

- Es necesario hacer notar que, cuando se quiere obtener una partición en  $K$  clases de un conjunto con  $n$  individuos, no tiene sentido examinar *todas* las posibles particiones del conjunto de individuos en  $K$  clases.
- En efecto, se está en presencia de un problema combinatorio muy complejo; sólo para efectos de ilustración, mencionemos que el número de particiones en 2 clases de un conjunto con 60 elementos es aproximadamente  $10^{18}$ , y para 100 elementos en 5 clases anda por  $10^{68}$ .



# Objetivo del Método K-means

- Así, el objetivo en el método de K-means es encontrar una partición  $P$  de  $\Omega$  y representantes de las clases, tales que  $W(P)$  sea mínima.

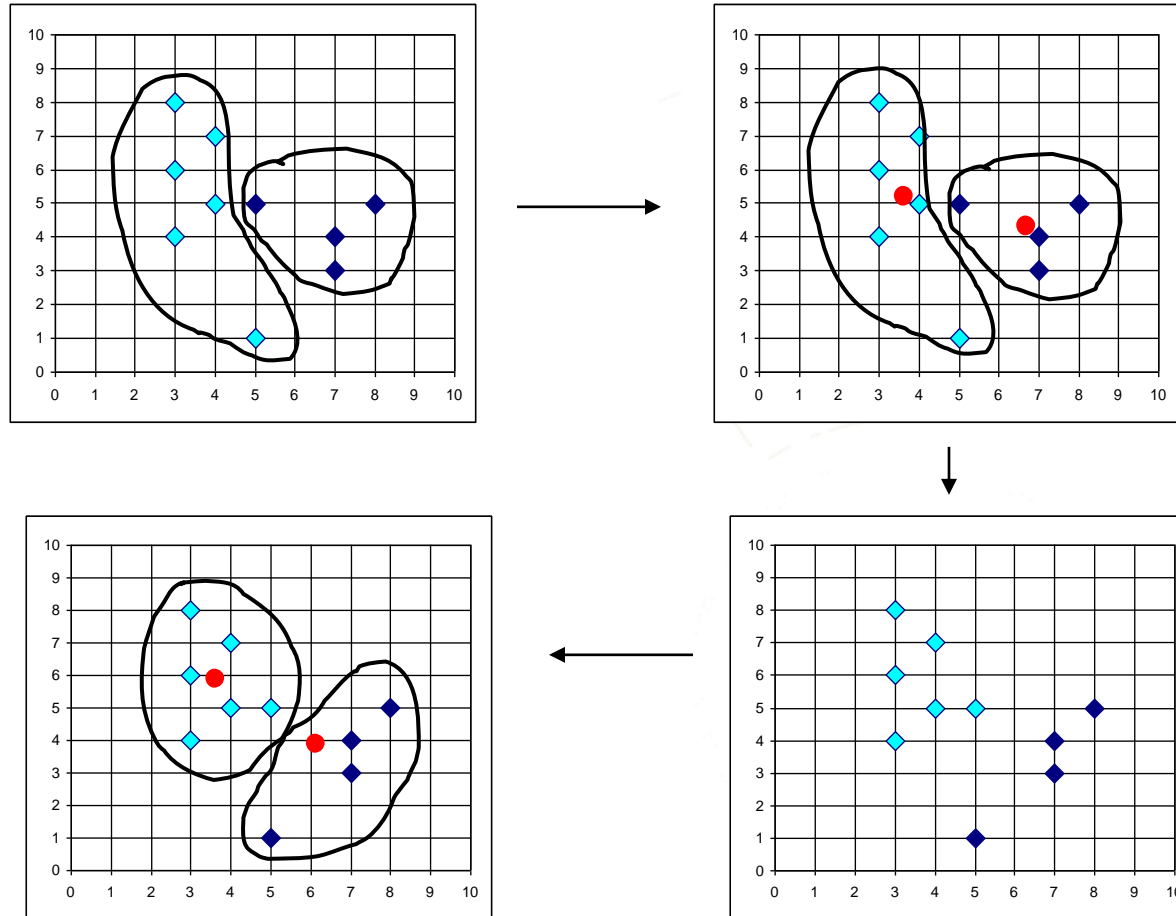
# Método de k-medias

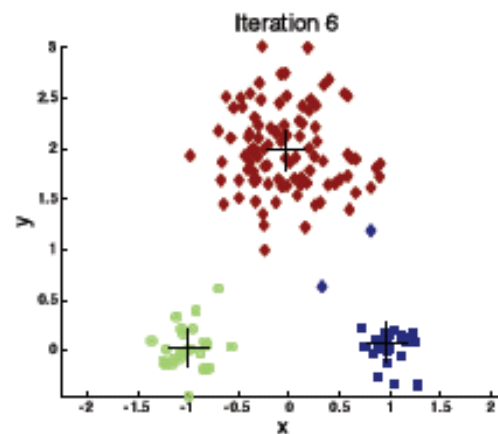
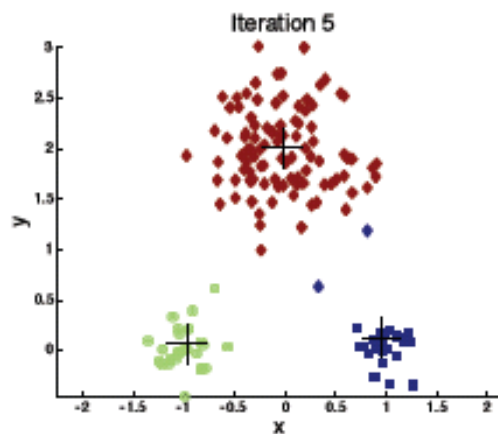
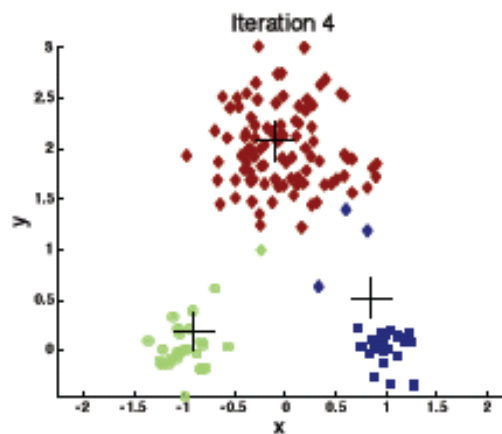
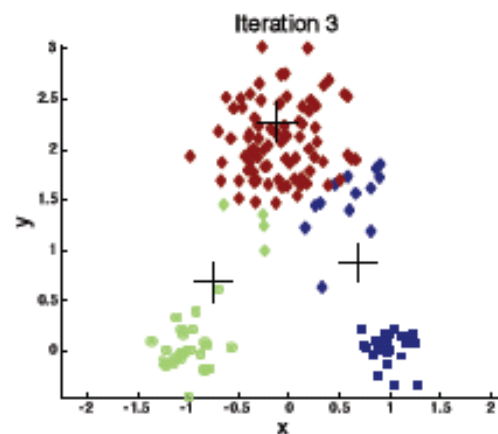
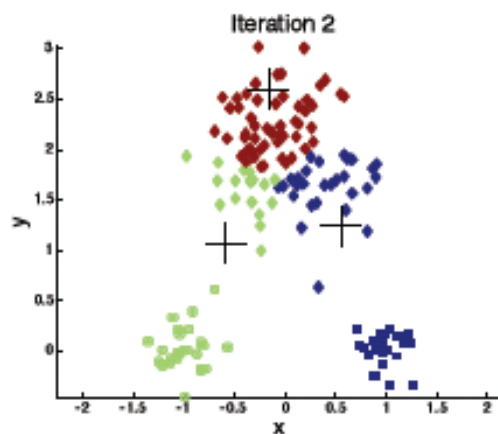
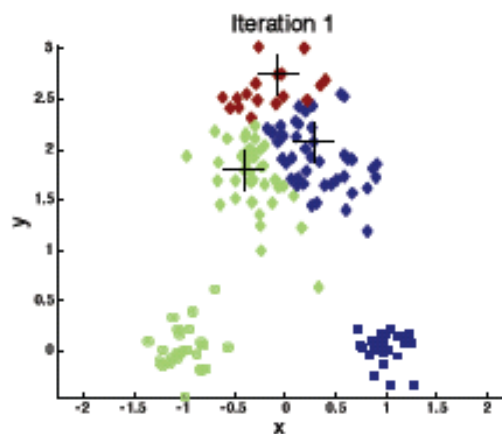
- Existe un poco de confusión en la literatura acerca del método de las k-medias, ya que hay dos métodos distintos que son llamados con el mismo nombre.
- Originalmente, Forgy propuso en 1965 un primer método de reasignación-recentraje que consiste básicamente en la iteración sucesiva, hasta obtener convergencia, de las dos operaciones siguientes:

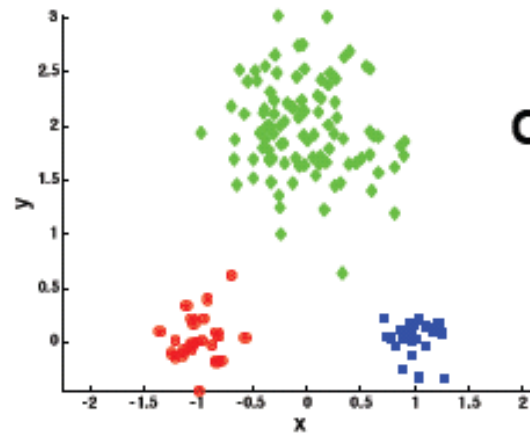


1. Representar una clase por su centro de gravedad, esto es, por su vector de promedios.
2. Asignar los objetos a la clase del centro de gravedad más cercano.

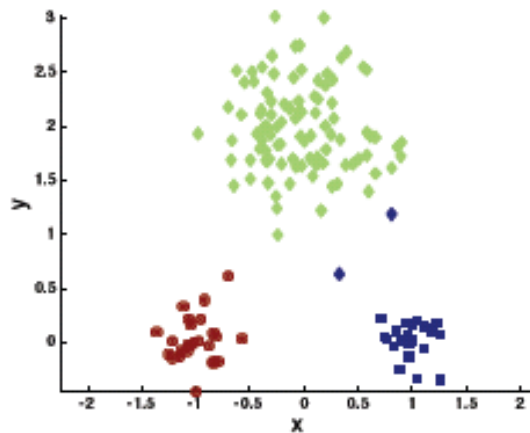
# The *K-Means* Clustering Method (nubes dinámicas)



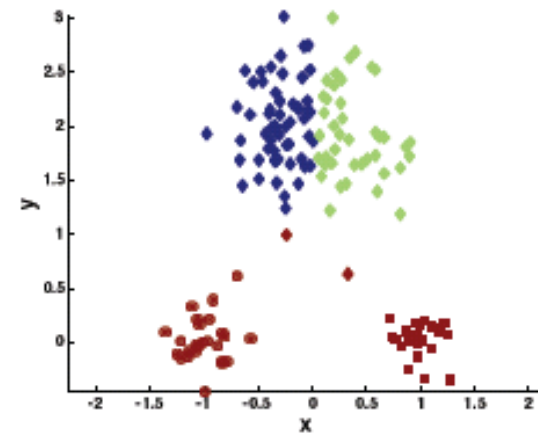




**Original Points**



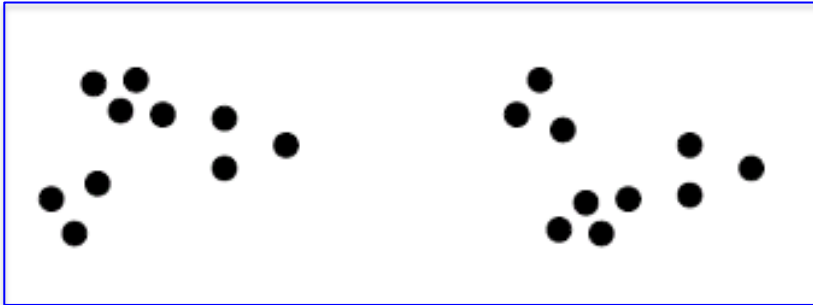
**Optimal Clustering**



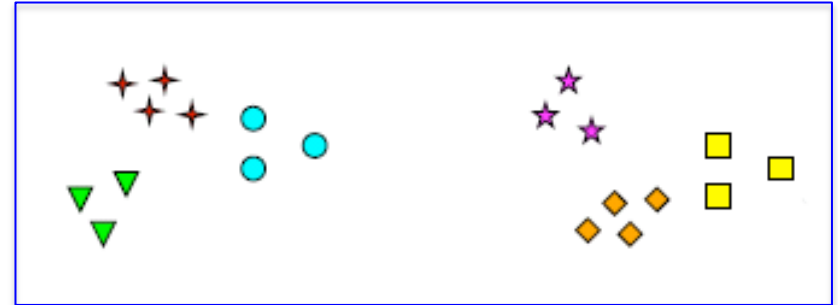
**Sub-optimal Clustering**



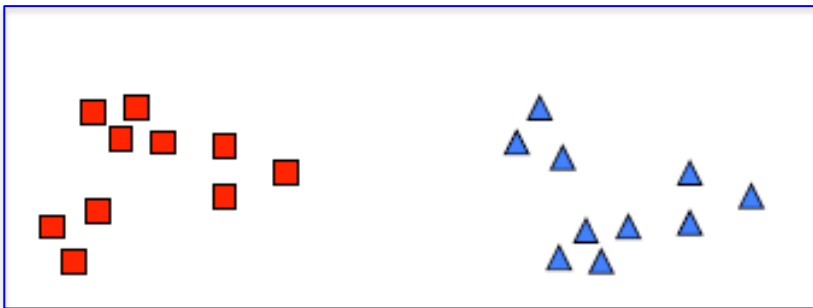
# ¿Cuántos clústeres?



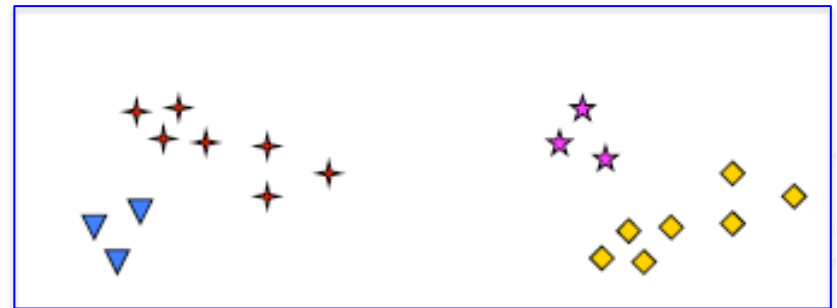
Datos originales



6 clústeres



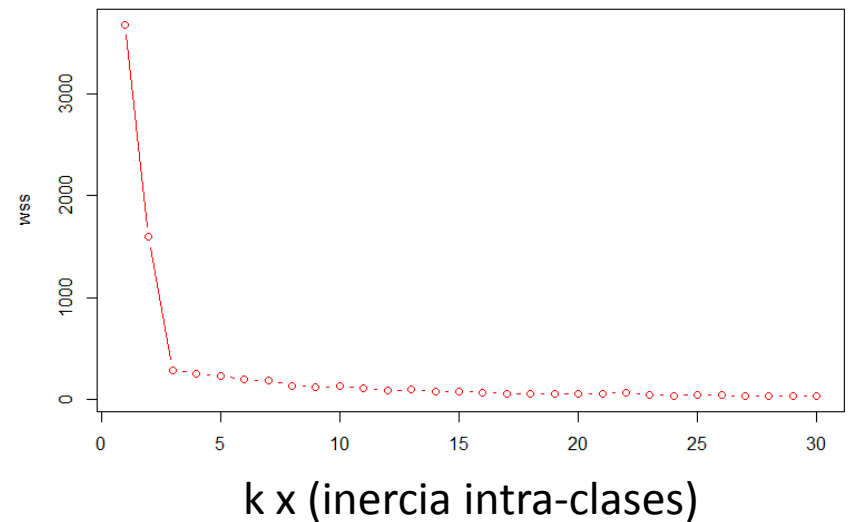
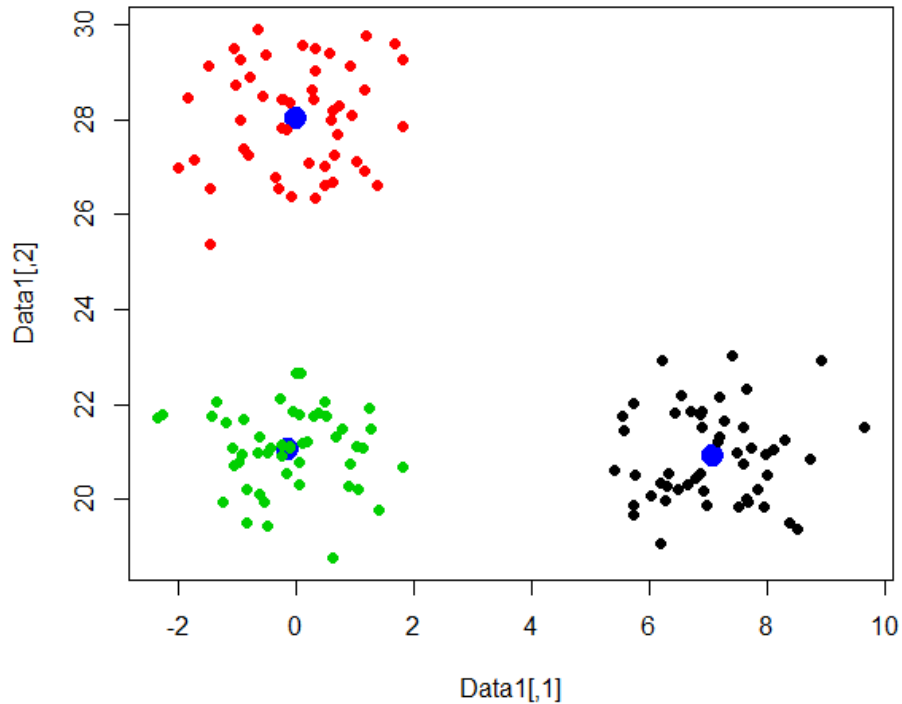
2 clústeres



4 clústeres



# ¿Cuántos clústeres?



El “codo” indica que  $k=3$  es la cantidad adecuada de clústeres



- McQueen propone un método muy similar, donde también se representan las clases por su centro de gravedad, y se examina cada individuo para asignarlo a la clase más cercana.
- La diferencia con el método de Forgy es que inmediatamente después de asignar un individuo a una clase, el centro de ésta es recalculado, mientras que Forgy primero hacía todas las asignaciones y luego recalculaba los centros.
- Variantes del método de Forgy son propuestas en Francia como Método de Nubes Dinámicas por E. Diday a partir en 1967.
- Es McQueen quien propone el nombre “k-means”, que se usa hasta la fecha, aún si estos métodos también reciben nombres como nubes dinámicas, centros móviles, o reasignación-recentraje.

# Método de Forgy

Denotaremos  $\Omega$  el conjunto de  $n$  individuos que queremos clasificar, todos dotados de pesos iguales  $1/n$ , y supondremos que están descritos por  $p$  variables cuantitativas  $x^1, x^2, \dots, x^p$ .

En el caso en que se está en presencia de variables cuantitativas, tiene sentido el cálculo de promedios y de distancias Euclídeas. Por lo tanto, también tiene sentido que cada clase esté representada por su centro de gravedad, esto es, por un individuo ficticio cuyas coordenadas son los valores promedio de las variables para los individuos pertenecientes a la clase. Este es el caso más simple y el usado más corrientemente. Generalmente, se usará la distancia Euclídea clásica en este contexto.





# Algoritmo: K-means

0. *Inicialización:* Escoger al azar  $K$  objetos de  $\Omega$ , que servirán como núcleos iniciales<sup>8</sup>. Esto es, escoger al azar  $\mathbf{g}_1, \dots, \mathbf{g}_K$  en  $\Omega$ ; sean

$$C_1 := \emptyset, \dots, C_K := \emptyset.$$

1. *Asignación:* Asignar cada objeto a la clase del centro de gravedad más cercano. Es decir, para todo  $i \in \Omega$  hacer:  
si  $d(\mathbf{x}_i, \mathbf{g}_{k^*}) \leq \{d(\mathbf{x}_i, \mathbf{g}_k) \text{ para todo } k = 1, \dots, K\}$  entonces asignar  $\mathbf{x}_i$  a la clase  $C_{k^*}$ ; si el mínimo se alcanza para dos clases diferentes entonces asignarlo a la clase de índice menor.
2. *Representación:* Calcular los centros de gravedad de la partición. Así, para todo  $k \in \{1, \dots, K\}$  hacer:  $\mathbf{g}_k := \frac{1}{|C_k|} \sum_{\mathbf{x}_i \in C_k} \mathbf{x}_i$ .  
Calcular el criterio  $W := \sum_{k=1}^K \sum_{i \in C_k} \|\mathbf{x}_i - \mathbf{g}_k\|^2$ .



# Algoritmo: K-means

3. *Control de parada:* Si la variación en el criterio  $W$  entre la iteración anterior y la presente es menor que un umbral dado, o si se sobrepasa el número máximo de iteraciones entonces detenerse, de lo contrario ir al paso 4.
4. *Preparación:* Poner  $C_1 := \emptyset, \dots, C_K := \emptyset$ ; ir al paso 1.

El resultado de la aplicación del método de k-medias, dependerá de la escogencia inicial de los núcleos. Por ello, se recomienda correr varias veces el método y escoger la mejor solución obtenida en esas corridas.



# K-Means Clustering Algorithm

---

## Algorithm 1 Basic K-means Algorithm.

---

- 1: Select  $K$  points as the initial centroids.
  - 2: **repeat**
  - 3:   Form  $K$  clusters by assigning all points to the closest centroid.
  - 4:   Recompute the centroid of each cluster.
  - 5: **until** The centroids don't change
- 



# Ejemplo de las notas escolares

Partición $P$	Número de veces obtenida	$W(P)$	$B(P)$
$C_1 = \{\text{Lucía, Andrés, Carlos, María}\}$ $C_2 = \{\text{Luis, Sonia}\}$ $C_3 = \{\text{Pedro, Inés, Ana, José}\}$	17 (68%)	0.75	4.97
$C_1 = \{\text{Lucía, Andrés, Carlos, María, Luis, Sonia}\}$ $C_2 = \{\text{Pedro, Inés}\}$ $C_3 = \{\text{Ana, José}\}$	3 (12%)	2.48	3.24
$C_1 = \{\text{Lucía, Andrés, Carlos, María, Luis, Sonia}\}$ $C_2 = \{\text{Inés, Ana, José}\}$ $C_3 = \{\text{Pedro}\}$	2 (8%)	2.52	3.20
$C_1 = \{\text{Lucía, Andrés, Carlos, María, Luis, Sonia}\}$ $C_2 = \{\text{Inés, Ana}\}$ $C_3 = \{\text{Pedro, José}\}$	1 (4%)	2.55	3.17
$C_1 = \{\text{Lucía, Andrés, Carlos, Luis, Sonia}\}$ $C_2 = \{\text{Pedro, Inés}\}$ $C_3 = \{\text{Ana, José, María}\}$	1 (4%)	2.72	3.00
$C_1 = \{\text{Lucía, Andrés, Carlos, María, Pedro, Inés, Ana, José}\}$ $C_2 = \{\text{Luis}\}$ $C_3 = \{\text{Sonia}\}$	1 (4%)	3.06	2.66



# *K-means en R*

## Description

Perform k-means clustering on a data matrix.

## Usage

```
kmeans(x, centers, iter.max = 10, nstart = 1,  
       algorithm = c("Hartigan-Wong", "Lloyd", "Forgy",  
                     "MacQueen"))  
## S3 method for class 'kmeans'  
fitted(object, method = c("centers", "classes"), ...)
```

## Arguments

- x** numeric matrix of data, or an object that can be coerced to such a matrix (such as a numeric vector or a data frame with all numeric columns).
- centers** either the number of clusters, say  $k$ , or a set of initial (distinct) cluster centres. If a number, a random set of (distinct) rows in **x** is chosen as the initial centres.
- iter.max** the maximum number of iterations allowed.
- nstart** if **centers** is a number, how many random sets should be chosen?
- algorithm** character: may be abbreviated.
- object** an R object of class "kmeans", typically the result of `ob <- kmeans(...)`.
- method** character: may be abbreviated. "centers" causes **fitted** to return cluster centers (one for each input point) and "classes" causes **fitted** to return a vector of class assignments.
- ...** not used.

# *K-means en R*

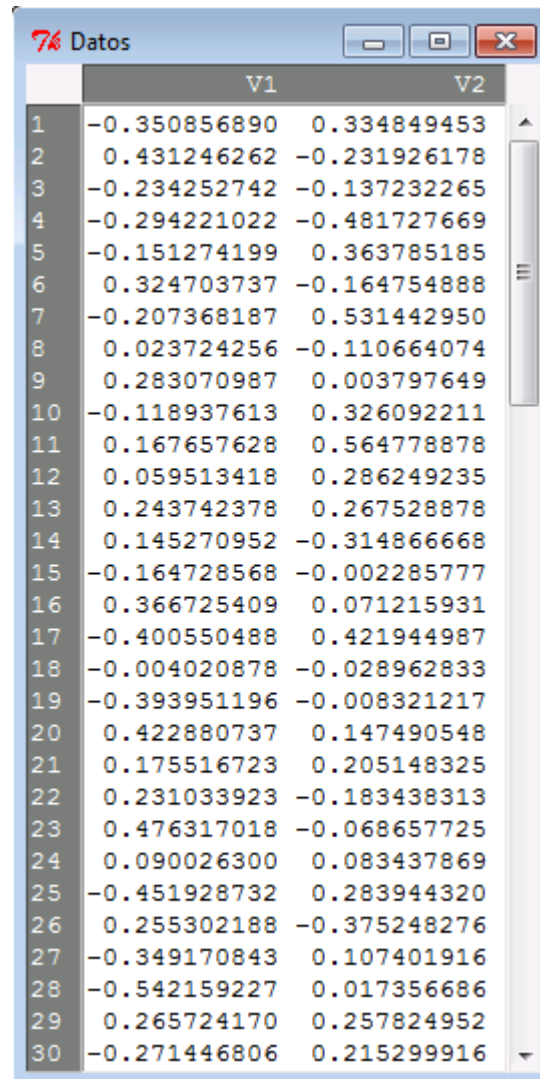
## Value

`kmeans` returns an object of class "kmeans" which has a `print` and a `fitted` method. It is a list with components:

<code>cluster</code>	A vector of integers (from <code>1:k</code> ) indicating the cluster to which each point is allocated.
<code>centers</code>	A matrix of cluster centres.
<code>totss</code>	The total sum of squares.
<code>withinss</code>	Vector of within-cluster sum of squares, one component per cluster.
<code>tot.withinss</code>	Total within-cluster sum of squares, i.e., <code>sum(withinss)</code> .
<code>betweenss</code>	The between-cluster sum of squares, i.e. <code>totss-tot.withinss</code> .
<code>size</code>	The number of points in each cluster.



# Ejemplo 2D: Datos

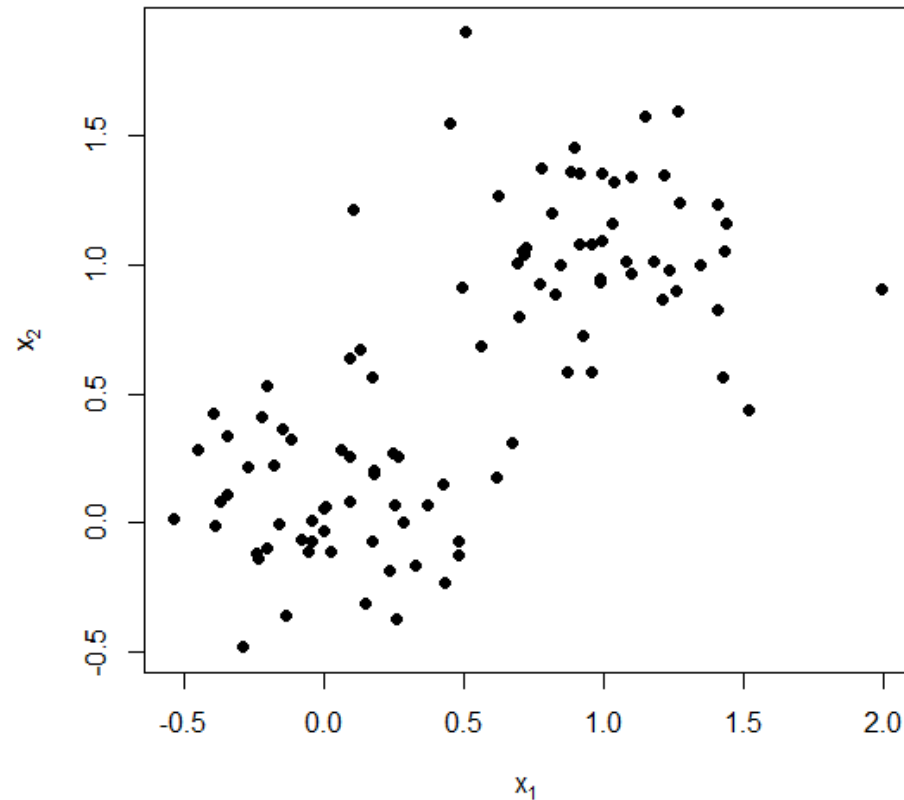


76 Datos

	V1	V2
1	-0.350856890	0.334849453
2	0.431246262	-0.231926178
3	-0.234252742	-0.137232265
4	-0.294221022	-0.481727669
5	-0.151274199	0.363785185
6	0.324703737	-0.164754888
7	-0.207368187	0.531442950
8	0.023724256	-0.110664074
9	0.283070987	0.003797649
10	-0.118937613	0.326092211
11	0.167657628	0.564778878
12	0.059513418	0.286249235
13	0.243742378	0.267528878
14	0.145270952	-0.314866668
15	-0.164728568	-0.002285777
16	0.366725409	0.071215931
17	-0.400550488	0.421944987
18	-0.004020878	-0.028962833
19	-0.393951196	-0.008321217
20	0.422880737	0.147490548
21	0.175516723	0.205148325
22	0.231033923	-0.183438313
23	0.476317018	-0.068657725
24	0.090026300	0.083437869
25	-0.451928732	0.283944320
26	0.255302188	-0.375248276
27	-0.349170843	0.107401916
28	-0.542159227	0.017356686
29	0.265724170	0.257824952
30	-0.271446806	0.215299916



# Ploteo de los Datos



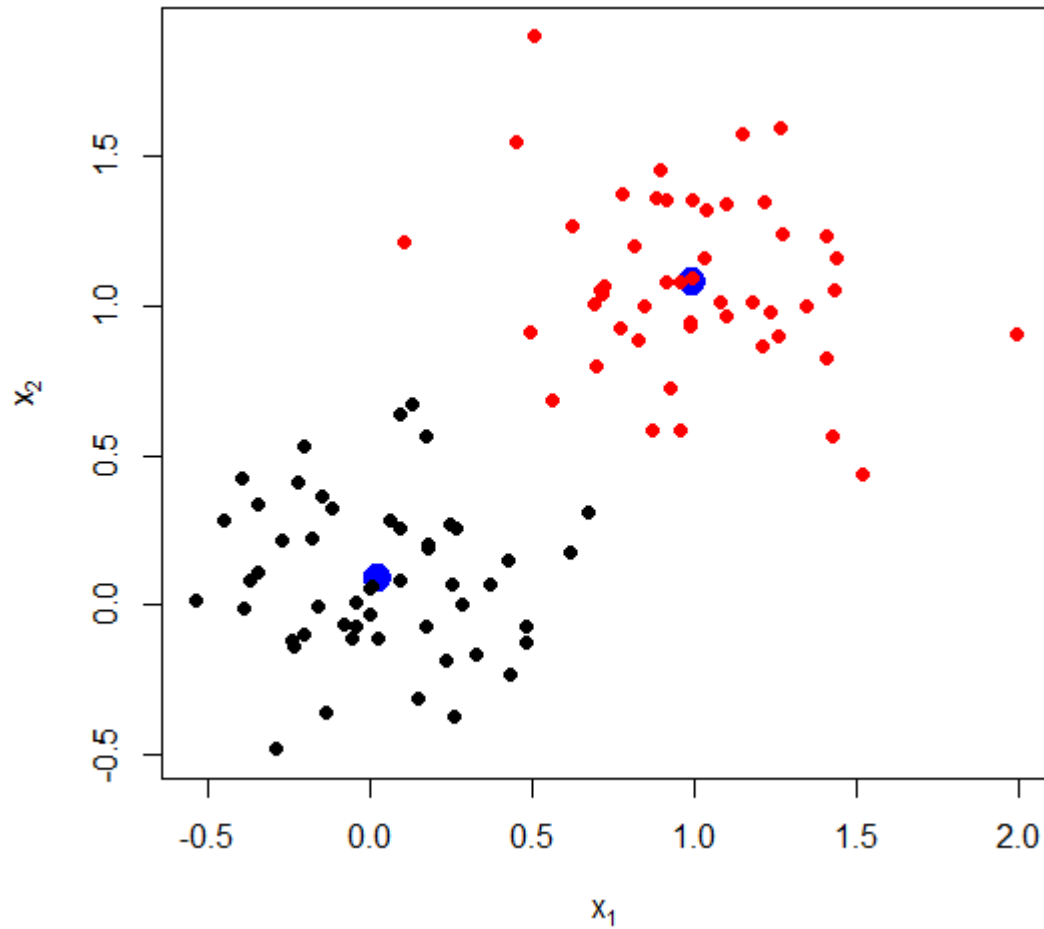


# K-means en R

```
setwd("C:/Users/Oldemar/Google Drive/Curso Minería Datos II -  
  Optativo/Datos")  
datos<-read.csv("Ej1kmeans.csv",sep = ";",header=F)  
datos  
plot(datos,pch=19,xlab=expression(x[1]),ylab=expression(x[2]))  
grupos<-kmeans(datos, 2)  
points(grupos$centers,pch=19,col="blue",cex=2)  
points(datos,col=grupos$cluster,pch=19)  
grupos$cluster  
grupos$centers
```



# K-means en R

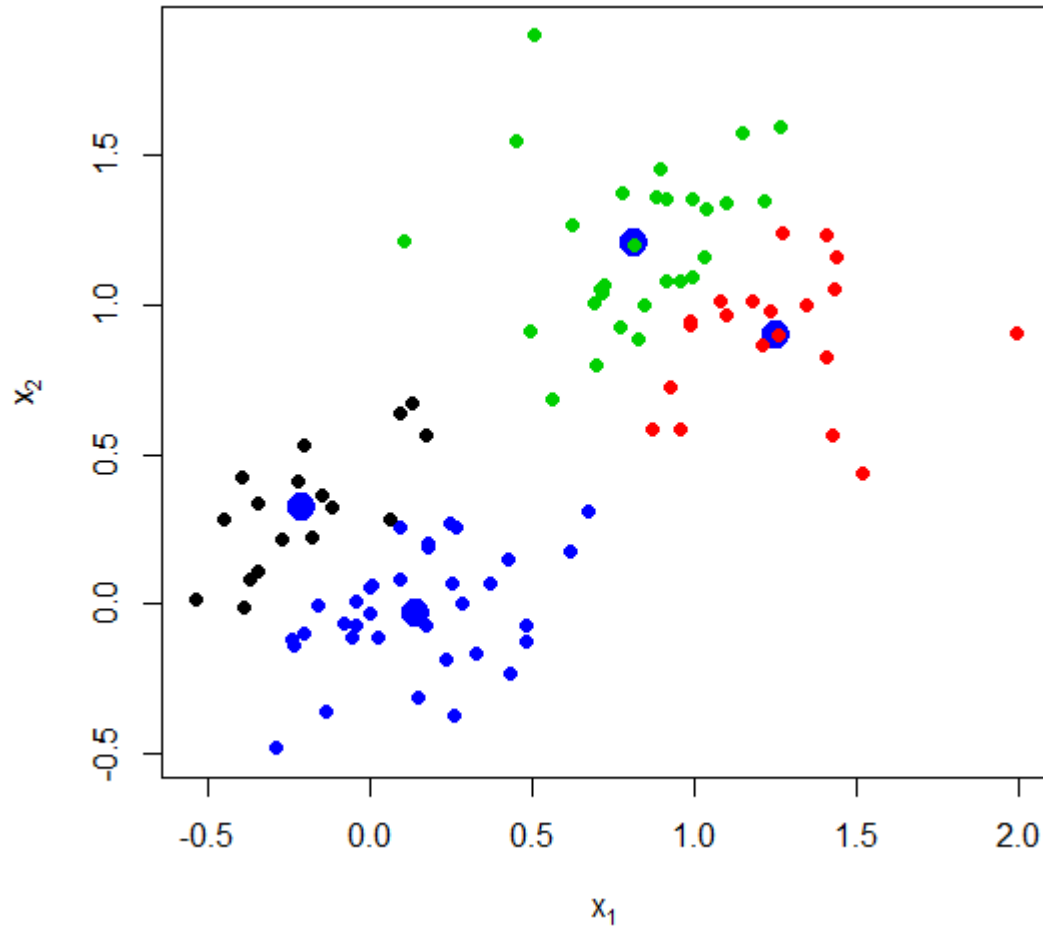


# K-means en R

```
setwd("C:/Users/Oldemar/Google Drive/Curso Minería Datos II -  
  Optativo/Datos")  
datos<-read.csv("Ej1kmeans.csv",sep = ";",header=F)  
datos  
plot(datos,pch=19,xlab=expression(x[1]),ylab=expression(x[2]))  
grupos<-kmeans(datos, 4)  
points(grupos$centers,pch=19,col="blue",cex=2)  
points(datos,col=grupos$cluster,pch=19)  
grupos$cluster  
grupos$centers
```



# K-means en R



# grupos\$cluster

```
> grupos$cluster
```

```
[1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
2 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2  
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
```



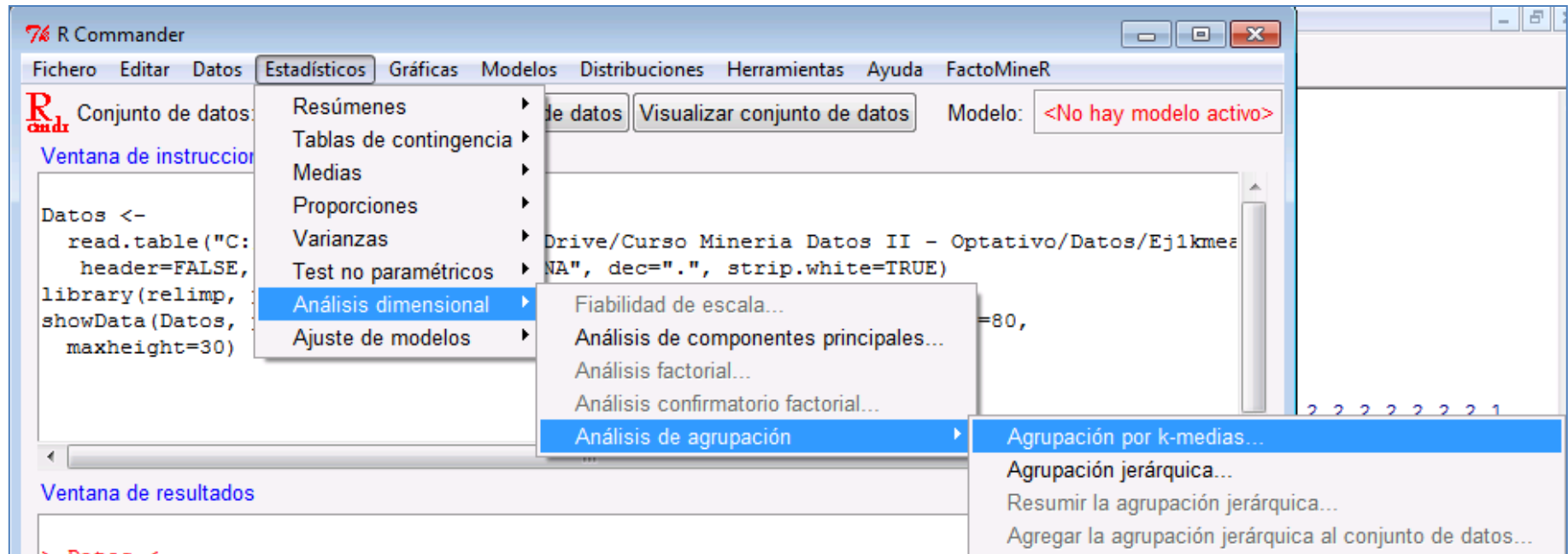
# grupos\$centers

```
> grupos$centers
```

	V1	V2
1	0.02169424	0.08865999
2	0.99128291	1.07898833



# En RComander



# En RComander

**Agrupación por k-medias**

Expresión de selección  
<todos los casos válido

Variables (seleccione una o más)

V1  
V2

Número de grupos: 2

Número de semillas iniciales: 10

Iteraciones máximas: 10

Imprimir resumen del grupo: ☒

Gráfica doble de grupos: ☒

Asignar grupos al conjunto de datos: ☒

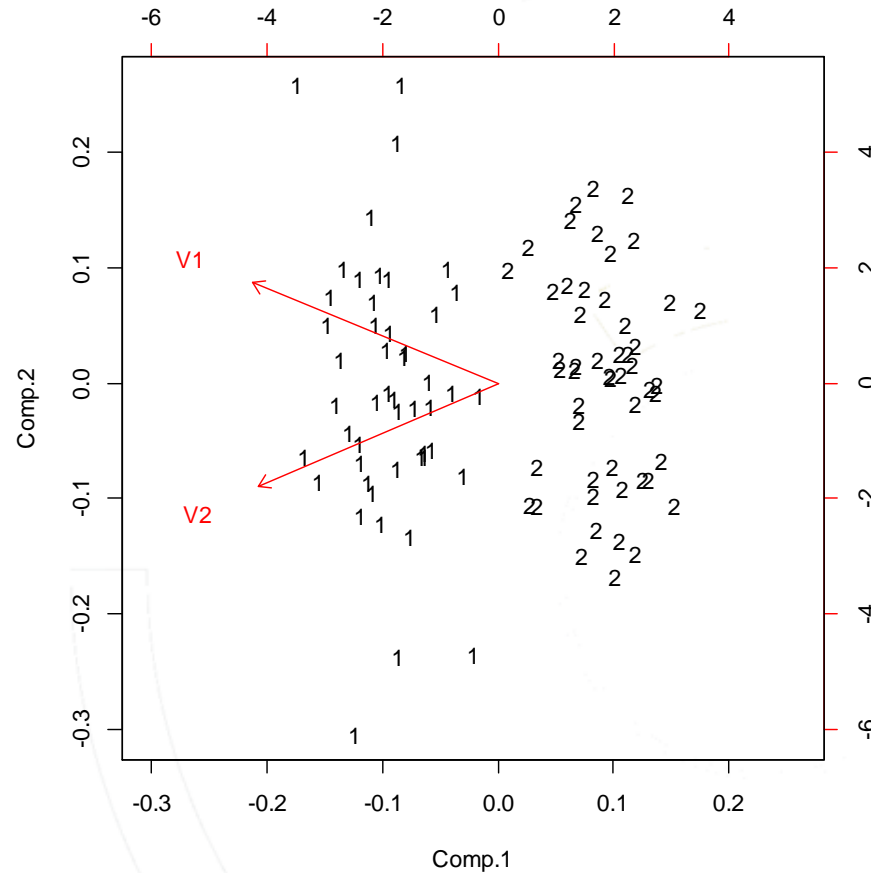
Variable de asignación: KMeans

Aceptar Cancelar Reiniciar Ayuda

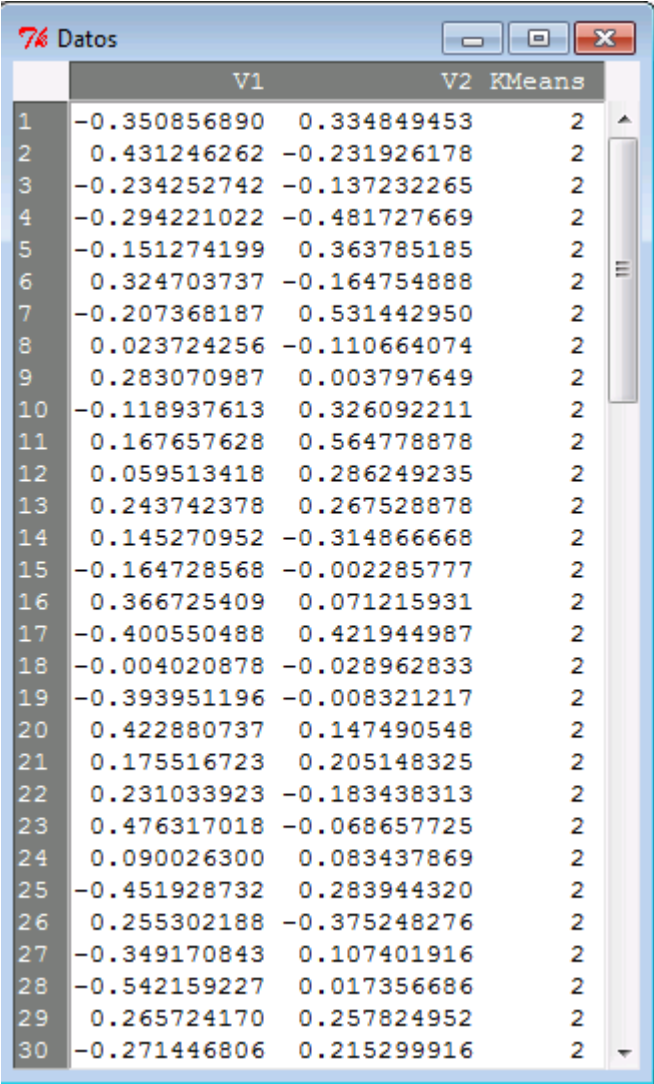




# En RComander



# En RComander



7% Datos

	V1	V2	KMeans	
1	-0.350856890	0.334849453	2	▲
2	0.431246262	-0.231926178	2	
3	-0.234252742	-0.137232265	2	
4	-0.294221022	-0.481727669	2	
5	-0.151274199	0.363785185	2	
6	0.324703737	-0.164754888	2	
7	-0.207368187	0.531442950	2	
8	0.023724256	-0.110664074	2	
9	0.283070987	0.003797649	2	
10	-0.118937613	0.326092211	2	
11	0.167657628	0.564778878	2	
12	0.059513418	0.286249235	2	
13	0.243742378	0.267528878	2	
14	0.145270952	-0.314866668	2	
15	-0.164728568	-0.002285777	2	
16	0.366725409	0.071215931	2	
17	-0.400550488	0.421944987	2	
18	-0.004020878	-0.028962833	2	
19	-0.393951196	-0.008321217	2	
20	0.422880737	0.147490548	2	
21	0.175516723	0.205148325	2	
22	0.231033923	-0.183438313	2	
23	0.476317018	-0.068657725	2	
24	0.090026300	0.083437869	2	
25	-0.451928732	0.283944320	2	
26	0.255302188	-0.375248276	2	
27	-0.349170843	0.107401916	2	
28	-0.542159227	0.017356686	2	
29	0.265724170	0.257824952	2	
30	-0.271446806	0.215299916	2	▼

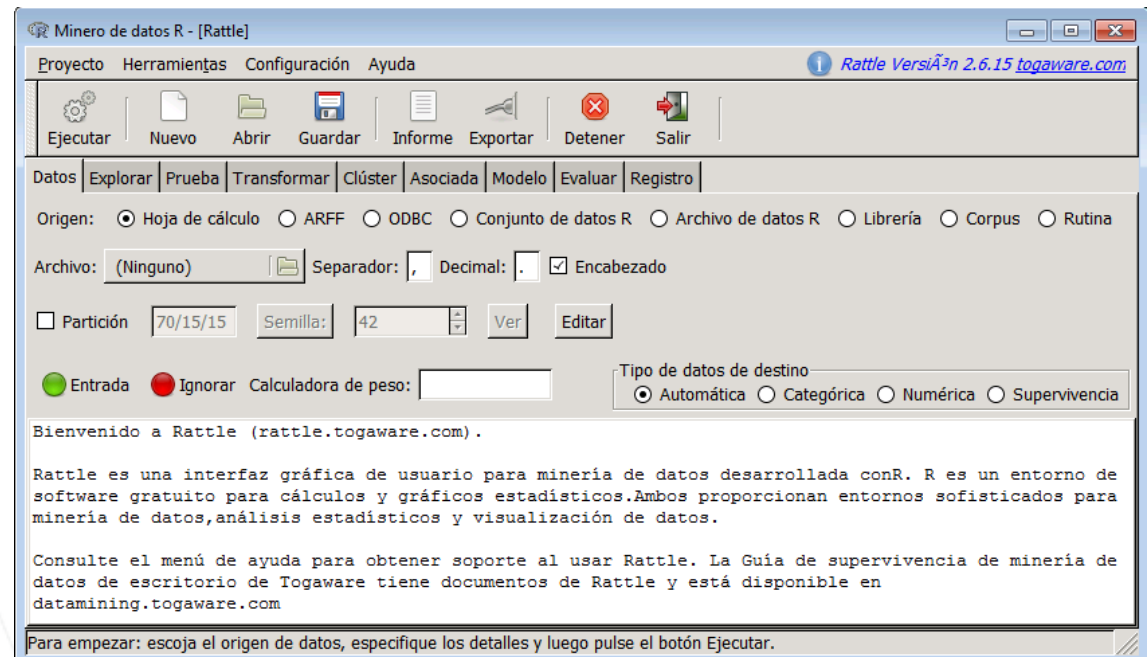


# ***Con Rattle***

- Para instalarlo:
  - `install.packages("rattle",dependencies =TRUE)`
- Para ejecutarlo:
  - `library(rattle)`
  - `rattle()`
- Sitio WEB:
  - <http://rattle.togaware.com/>



Dr. Graham Williams is the author of the Rattle data mining software and Adjunct Professor, University of Canberra and Australian National University.



# Interfaz de Rattle

The screenshot shows the Rattle data mining software interface. The title bar reads "Minero de datos R - [Rattle]". The menu bar includes "Proyecto", "Herramientas", "Configuración", and "Ayuda". A status bar at the top right indicates "Rattle Versión 2.6.15 [togaware.com](http://togaware.com)".

The main toolbar contains icons for "Ejecutar" (Execute), "Nuevo" (New), "Abrir" (Open), "Guardar" (Save), "Informe" (Report), "Exportar" (Export), "Detener" (Stop), and "Salir" (Exit).

Below the toolbar is a tabbed interface with tabs for "Datos", "Explorar", "Prueba", "Transformar", "Clúster", "Asociada", "Modelo", "Evaluar", and "Registro". The "Datos" tab is currently selected.

Under the "Datos" tab, the "Origen:" (Origin) section has radio buttons for "Hoja de cálculo" (selected), "ARFF", "ODBC", "Conjunto de datos R", "Archivo de datos R", "Librería", "Corpus", and "Rutina".

The "Archivo:" (File) section shows "(Ninguno)" (None) in the file name field, a folder icon, a "Separador:" (Delimiter) dropdown set to ",", a "Decimal:" dropdown set to ".", and a checked "Encabezado" (Header) checkbox.

The "Partición" (Partition) section has an unchecked checkbox, a "70/15/15" split field, a "Semilla:" (Seed) field with the value "42", and "Ver" (View) and "Editar" (Edit) buttons.

The "Entrada" (Input) section has a green circle icon, a red circle icon labeled "Ignorar", and a "Calculadora de peso:" (Weight calculator) field.

The "Tipo de datos de destino" (Destination data type) section has radio buttons for "Automática" (selected), "Categorica", "Numerica", and "Supervivencia".

The main text area contains the following text:

Bienvenido a Rattle ([rattle.togaware.com](http://rattle.togaware.com)).

Rattle es una interfaz gráfica de usuario para minería de datos desarrollada con R. R es un entorno de software gratuito para cálculos y gráficos estadísticos. Ambos proporcionan entornos sofisticados para minería de datos, análisis estadísticos y visualización de datos.

Consulte el menú de ayuda para obtener soporte al usar Rattle. La Guía de supervivencia de minería de datos de escritorio de Togaware tiene documentos de Rattle y está disponible en [datamining.togaware.com](http://datamining.togaware.com)

Para empezar: escoja el origen de datos, especifique los detalles y luego pulse el botón Ejecutar.

# Cargando los Datos

Minero de datos R - [Rattle (Ej1kmeans.csv)]



Proyecto Herramientas Configuración Ayuda

Rattle VersiÃ³n 2.6.15 [togaware.com](http://togaware.com)

Ejecutar Nuevo Abrir Guardar Informe Exportar Detener Salir

Datos Explorar Prueba Transformar Clúster Asociada Modelo Evaluar Registro

Origen: ☒ Hoja de cálculo ☐ ARFF ☐ ODBC ☐ Conjunto de datos R ☐ Archivo de datos R ☐ Librería ☐ Corpus ☐ Rutina

Archivo:  Ej1kmeans.csv  Separador: ; Decimal: . ☐ Encabezado

☒ Partición 70/15/15 Semilla: 42 Ver Editar

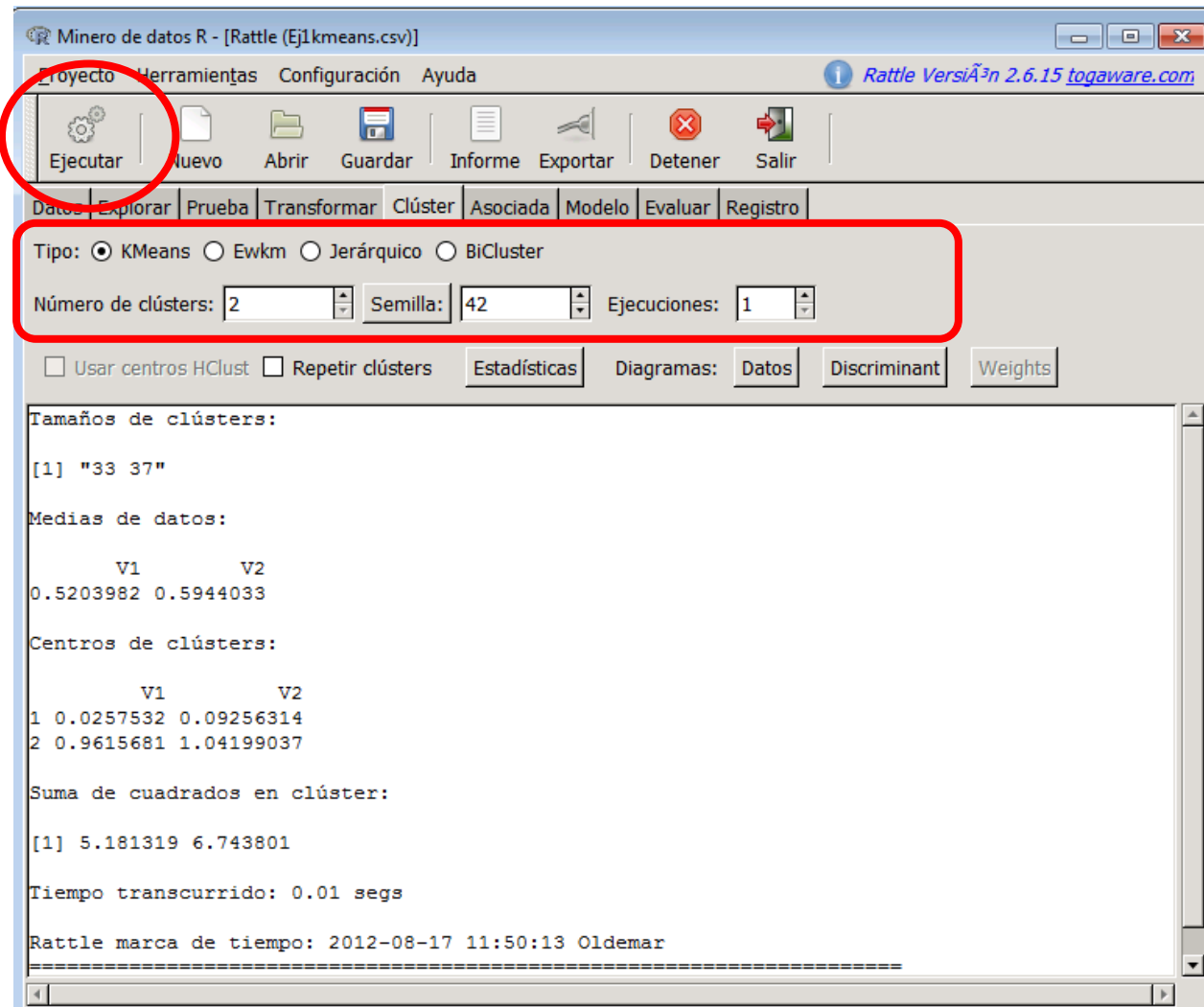
☒ Entrada ☒ Ignorar Calculadora de peso:

Tipo de datos de destino  
☒ Automática ☐ Categórica ☐ Numérica ☐ Supervivencia

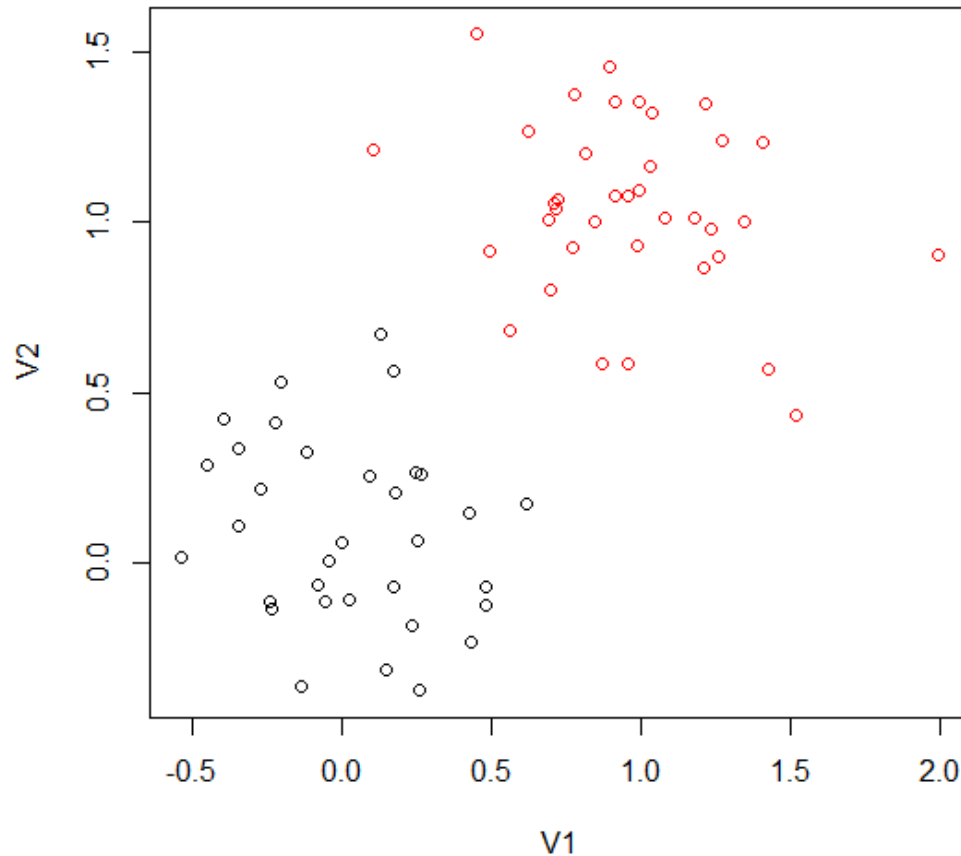
No.	Variable	Tipo de datos	Entrada	Destino	Riesgo	Ident	Ignorar	Weight	Comentario
1	V1	Numérica	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Única: 100
2	V2	Numérica	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Única: 100

Se han anotado los roles. Observaciones 100 y variables de entrada 2. No hay destino y por eso no hay modelo de predicción ni muestreo.

# Generando el modelo k-means



# Opción “Datos”

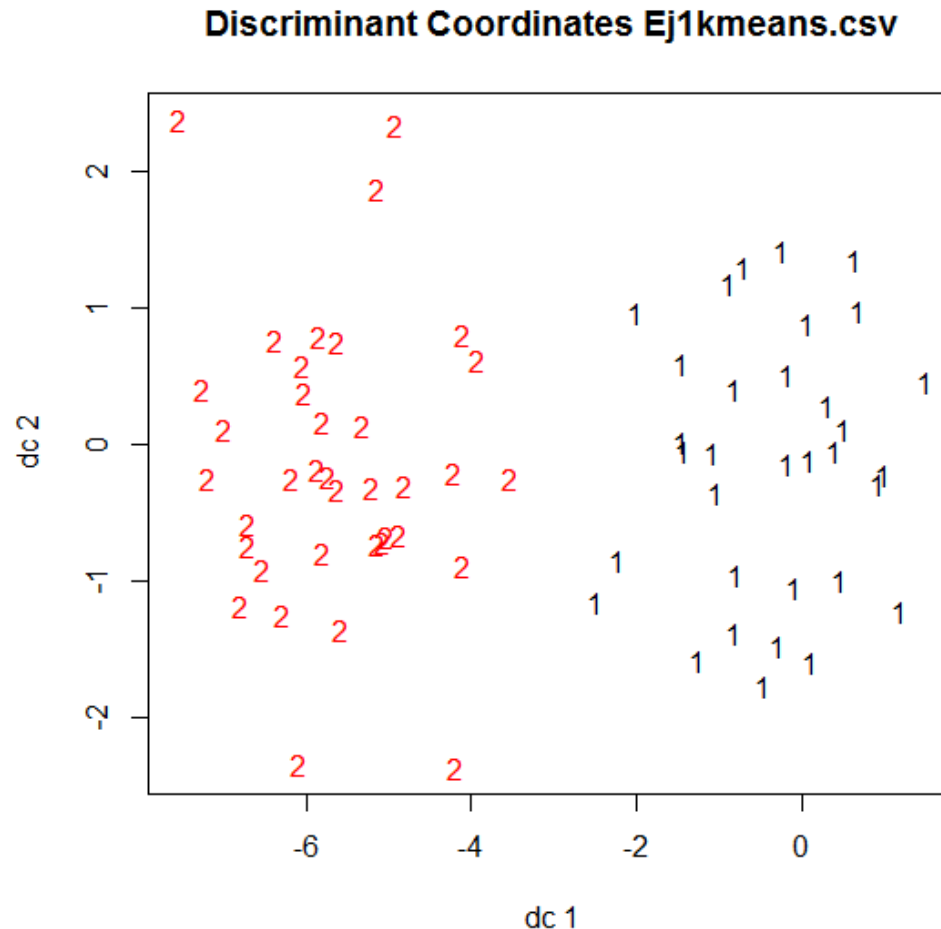


Rattle 2012-ago-17 11:55:09 Oldemar





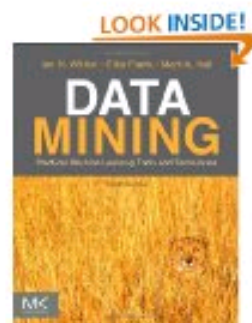
# Opción “Discriminat”



Rattle 2012-ago-17 11:56:54 Oldemar

# Algoritmos de Recomendación

## Recommendations for You in Books



Data Mining: Practical Machine...

Ian H. Witten, Eibe Frank, Mark A. Hall

Paperback

~~\$69.95~~ **\$42.12**

[Fix this recommendation](#)



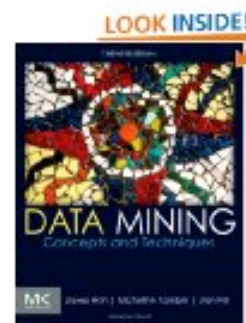
Multivariate Data Analysis (7th Edition)

Joseph F. Hair, William C. Black, ...

Hardcover

~~\$215.24~~ **\$215.24**

[Fix this recommendation](#)



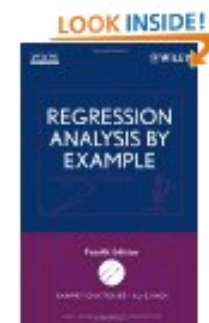
Data Mining: Concepts and Techniques...

Jiawei Han, Micheline Kamber, Jian Pei

Hardcover

~~\$74.95~~ **\$50.03**

[Fix this recommendation](#)



Regression Analysis by Example

> Bertram Price, Samprit Chatterjee

Hardcover

~~\$152.00~~ **\$110.24**

[Fix this recommendation](#)



**oldemar rodríguez**  
CONSULTOR en MINERÍA DE DATOS

## Recommendations for You in Clothing



adidas Boys 2-7 Revolution  
Basic Pant

~~\$19.99~~ - ~~\$21.00~~

[Fix this recommendation](#)



Sesame Street Ernie & Bert  
Homespun...

~~\$15.95~~

[Fix this recommendation](#)



adidas Infant Boys Core  
tricot Set

~~\$23.92~~ - ~~\$25.20~~

[Fix this recommendation](#)



Sesame Street Elmo &  
Friends Varsity...

~~\$14.95~~

[Fix this recommendation](#)

[➤ See more recommendations](#)

## Recommendations for You in Toys & Games



Cars 2 Collector  
Submariner...

Mattel

~~\$17.99~~

[Fix this recommendation](#)



Cars 2 Radiator Springs  
Playtown

Mattel

~~\$43.99~~ ~~\$25.82~~

[Fix this recommendation](#)



Disney Cars 2 Lightning  
McQueen

~~\$7.99~~ ~~\$2.76~~

[Fix this recommendation](#)



Disney / Pixar CARS 2  
Movie 155 Die...

~~\$7.99~~ ~~\$3.99~~

[Fix this recommendation](#)



**oldemar rodríguez**  
CONSULTOR en MINERÍA DE DATOS

## Recommendations for You in Electronics



Epson Duet 80-Inch Dual  
Aspect Ratio...

[Fix this recommendation](#)



HP Laserjet 125A Dual  
Black Cartridge...

[Fix this recommendation](#)



Link Depot HDMI to HDMI  
Cable (25 feet)

[Fix this recommendation](#)



Deluxe 425 Project-o Stand  
Projection...

Da-Lite

[Fix this recommendation](#)

[> See more recommendations](#)

## Recommendations for You in Musical Instruments



1st Note Metal Kazoo  
Woodstock

~~\$6.95~~ **\$4.89**

[Fix this recommendation](#)



Trophy Snoopy Jaws Harp  
**\$6.23**

[Fix this recommendation](#)



Acme Siren Whistle  
~~\$75.95~~ **\$31.73**

[Fix this recommendation](#)



Trophy 3470 Musical  
Spoons

~~\$9.95~~ **\$6.69**

[Fix this recommendation](#)



**oldemar rod  guez**  
CONSULTOR en MINER  A DE DATOS

## Customer Reviews



3.4 out of 5 stars

5 star		10
4 star		10
3 star		4
2 star		5
1 star		6

[See all 35 customer reviews](#)

## RODRIGUEZ, create your review

(If you're not RODRIGUEZ Oldemar, [click here.](#))

### Start here

Set your pen name: ([learn more](#))

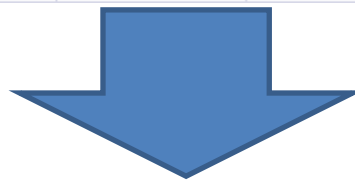
- 1 Are you over 13? ☐ Yes ☐ No
- 2 How much fun is this toy? ☆☆☆☆☆  
How would you rate this toy's educational value? ☆☆☆☆☆  
How durable is this toy? ☆☆☆☆☆  
How do you rate this toy overall? ☆☆☆☆☆
- 3 Please enter a title for your review:



**oldemar rodríguez**  
CONSULTOR en MINERÍA DE DATOS

# Tabla con los promedios de evaluación de 100 personas que adquirieron los mismos productos o muy similares

	Velocidad Entrega	Precio	Durabilidad	Imagen Producto	Valor Educativo	Servicio Retorno	Tamano Paquete	Calidad Producto	Numero Estrellas
Adam	2,05	0,3	3,45	2,35	2,4	2,3	2,6	2,1	1,7
Anna	0,9	1,5	3,15	3,3	2,5	4	4,2	2,15	2,8
Bernard	1,7	2,6	2,85	3	4,3	2,7	4,1	2,6	3,3
Edward	1,35	0,5	3,55	2,95	1,8	2,3	3,9	1,95	1,7
Emilia	3	0,45	4,8	3,9	3,4	4,6	2,25	3,4	4,3
Fabian	0,95	1,65	3,95	2,4	2,6	1,9	4,85	2,2	3
Philip	2,3	1,2	4,75	3,3	3,5	4,5	3,8	2,9	3,1
Frank	0,65	2,1	3,1	2,55	2,8	2,2	3,45	2,15	2,9
Xavier	2,75	0,8	4,7	2,35	3,5	3	3,8	2,7	4,8
Gabriel	2	1,75	3,25	3	3,7	3,2	4,35	2,7	3,9
Marisol	1,2	0,8	4,4	2,4	2	2,8	2,9	2,15	1,7
Henry	1,95	1,1	4,55	2,3	3	2,5	4,15	2,5	3,2
Irene	1,4	0,7	4,05	1,9	2,1	1,4	3,3	2,2	2,4
Isabelle	1,85	0,75	4,3	2,85	2,7	3,7	3,35	2,5	2,3
Isidore	2,35	0,65	4,95	3,35	3	2,6	3,4	2,95	3,9
Joseph	1,7	1	4,85	2,35	2,7	1,7	2,4	2,35	3,4



# K-means en R

```
setwd("C:/Users/Oldemar/Google Drive/Curso Minería Datos II -  
  Optativo/Datos")  
datos <- read.csv("C:/Users/Oldemar/Google Drive/Curso Minería Datos II -  
  Optativo/Datos/EjemploAlgoritmosRecomendación.csv",header=TRUE,  
  sep=";", dec=",", row.names=1)  
datos  
grupos<-kmeans(datos, 3)  
grupos$cluster  
grupos$centers
```

# grupos\$cluster

>grupos\$cluster

Adam	Anna	Bernard	Edward	Emilia	Fabian
3	1	1	3	2	3
Philip	Frank	Xavier	Gabriel	Marisol	Henry
1	3	2	1	3	2
Irene	Isabelle	Isidore	Joseph	Eugene	Eugenia
3	3	2	2	1	2
Eunice	Eva	Evdokia	Fedir	Felix	Fialka
2	2	3	3	1	3

.....



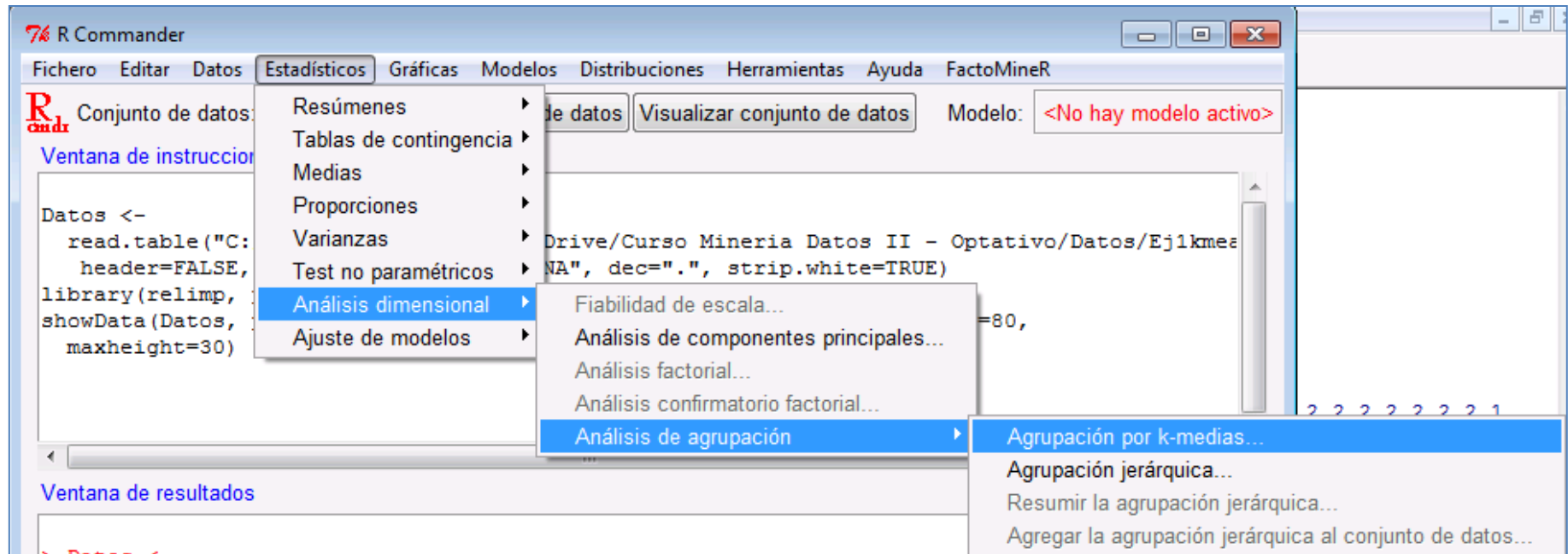


# grupos\$centers

> grupos\$centers

	Velocidad.Entrega	Precio	Durabilidad	Imagen.Producto	Valor.Educativo	Servicio.Retorno	Tamano.Paquete
1	1.629630	1.8703704	3.581481	3.072222	3.485185		
	3.225926	3.992593					
2	2.351316	0.8368421	4.542105	2.496053	3.192105		
	2.484211	2.906579					
3	1.211429	1.0257143	3.582857	2.417143	2.177143		
	2.428571	3.722857					
	Calidad.Producto	Numero.Estrellas					
1	2.522222	3.422222					
2	2.659211	3.757895					
3	1.982857	2.165714					

# En RComander



# En RComander

**Agrupación por k-medias**

Expresión de selección  
<todos los casos válido

Variables (seleccione una o más)

- Servicio.Retorno
- Tamano.Paquete
- Valor.Educativo
- Velocidad.Entrega

Número de grupos: 3

Número de semillas iniciales: 10

Iteraciones máximas: 10

Imprimir resumen del grupo: ☒

Gráfica doble de grupos: ☒

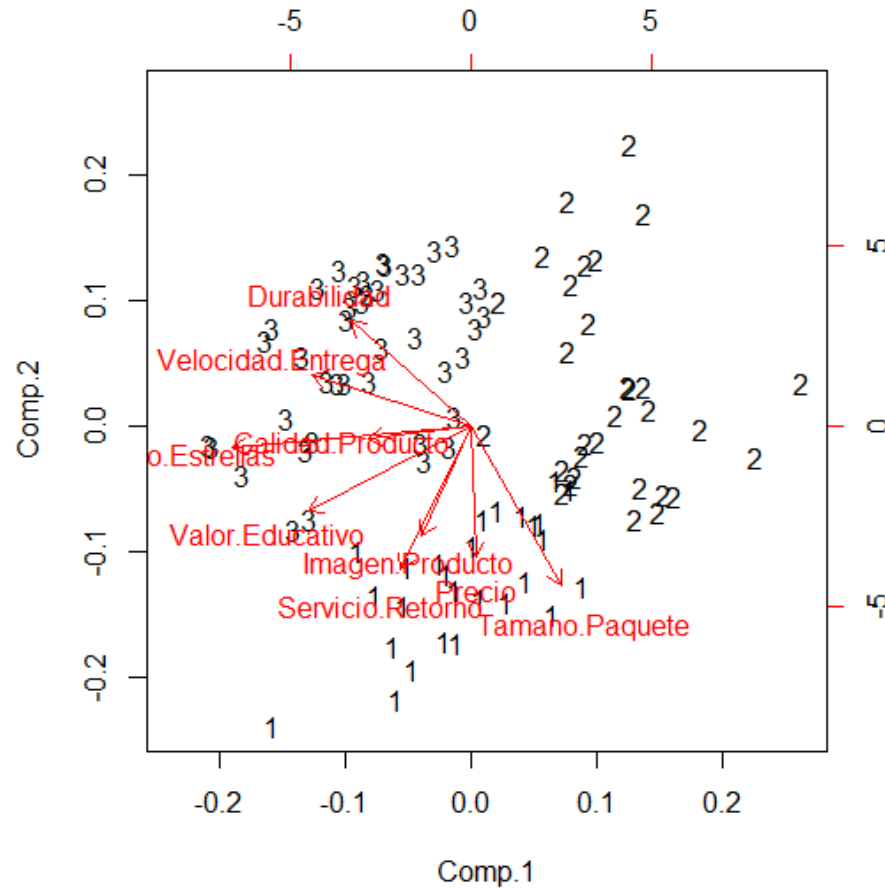
Asignar grupos al conjunto de datos: ☒

Variable de asignación: KMeans

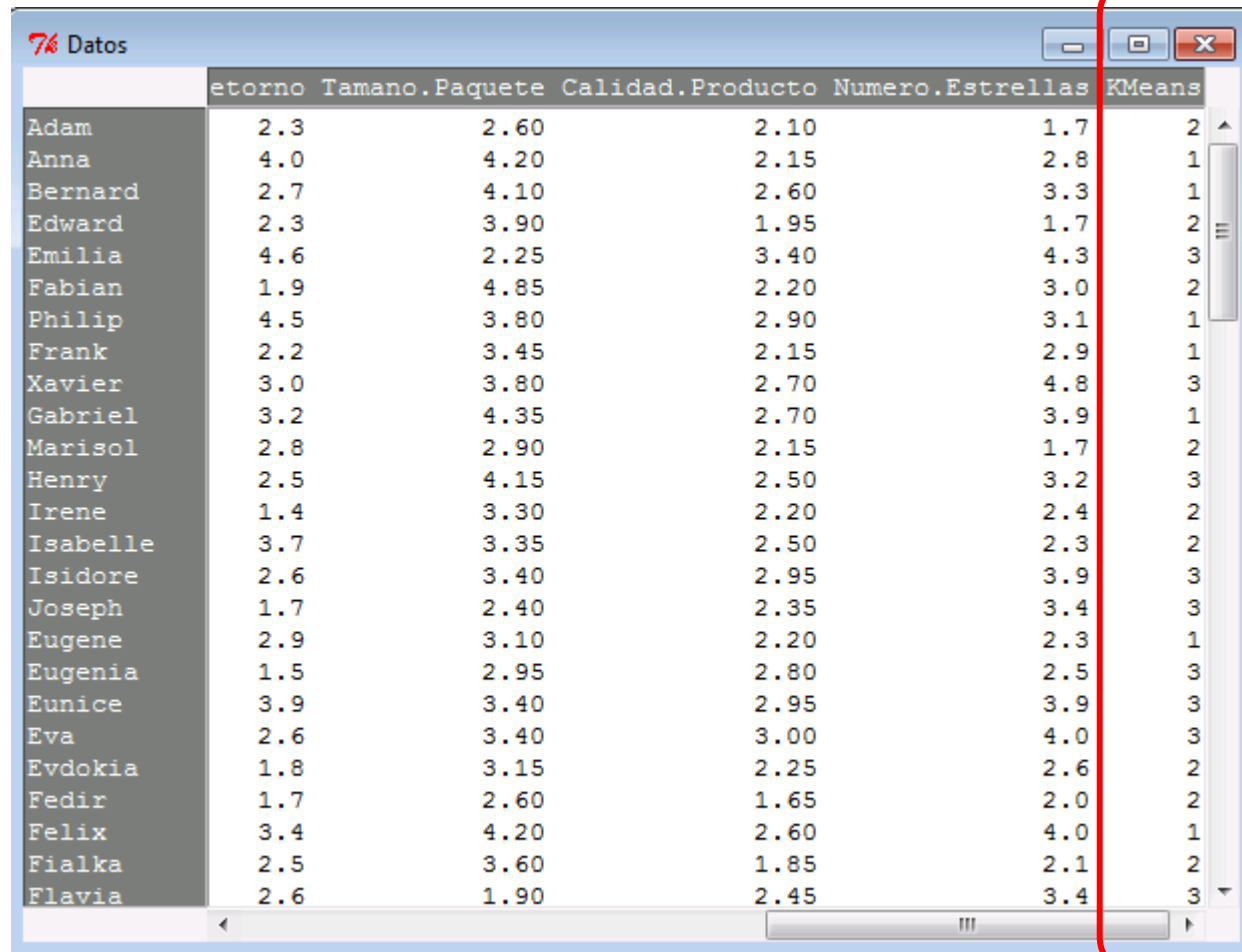
Aceptar Cancelar Reiniciar Ayuda



# En RComander



# En RComander



76 Datos

	etorno	Tamano.Paquete	Calidad.Producto	Numero.Estrellas	KMeans
Adam	2.3	2.60	2.10	1.7	2
Anna	4.0	4.20	2.15	2.8	1
Bernard	2.7	4.10	2.60	3.3	1
Edward	2.3	3.90	1.95	1.7	2
Emilia	4.6	2.25	3.40	4.3	3
Fabian	1.9	4.85	2.20	3.0	2
Philip	4.5	3.80	2.90	3.1	1
Frank	2.2	3.45	2.15	2.9	1
Xavier	3.0	3.80	2.70	4.8	3
Gabriel	3.2	4.35	2.70	3.9	1
Marisol	2.8	2.90	2.15	1.7	2
Henry	2.5	4.15	2.50	3.2	3
Irene	1.4	3.30	2.20	2.4	2
Isabelle	3.7	3.35	2.50	2.3	2
Isidore	2.6	3.40	2.95	3.9	3
Joseph	1.7	2.40	2.35	3.4	3
Eugene	2.9	3.10	2.20	2.3	1
Eugenia	1.5	2.95	2.80	2.5	3
Eunice	3.9	3.40	2.95	3.9	3
Eva	2.6	3.40	3.00	4.0	3
Evdokia	1.8	3.15	2.25	2.6	2
Fedir	1.7	2.60	1.65	2.0	2
Felix	3.4	4.20	2.60	4.0	1
Fialka	2.5	3.60	1.85	2.1	2
Flavia	2.6	1.90	2.45	3.4	3



# ***Con Rattle***

- Para instalarlo:
  - `install.packages("rattle",dependencies =TRUE)`
- Para ejecutarlo:
  - `library(rattle)`
  - `rattle()`
- Sitio WEB:
  - <http://rattle.togaware.com/>



# Cargando los Datos

Minero de datos R - [Rattle (EjemploAlgoritmosRecomendación.csv)]

Proyecto Herramientas Configuración Ayuda

Rattle Versión 2.6.15 [togaware.com](http://togaware.com)

Ejecutar Nuevo Abrir Guardar Informe Exportar Detener Salir

Datos Explorar Prueba Transformar Clúster Asociada Modelo Evaluar Registro

Origen: ☒ Hoja de cálculo ☐ ARFF ☐ ODBC ☐ Conjunto de datos R ☐ Archivo de datos R ☐ Librería ☐ Corpus ☐ Rutina

Archivo:  Separador: ; Decimal: , ☒ Encabezado

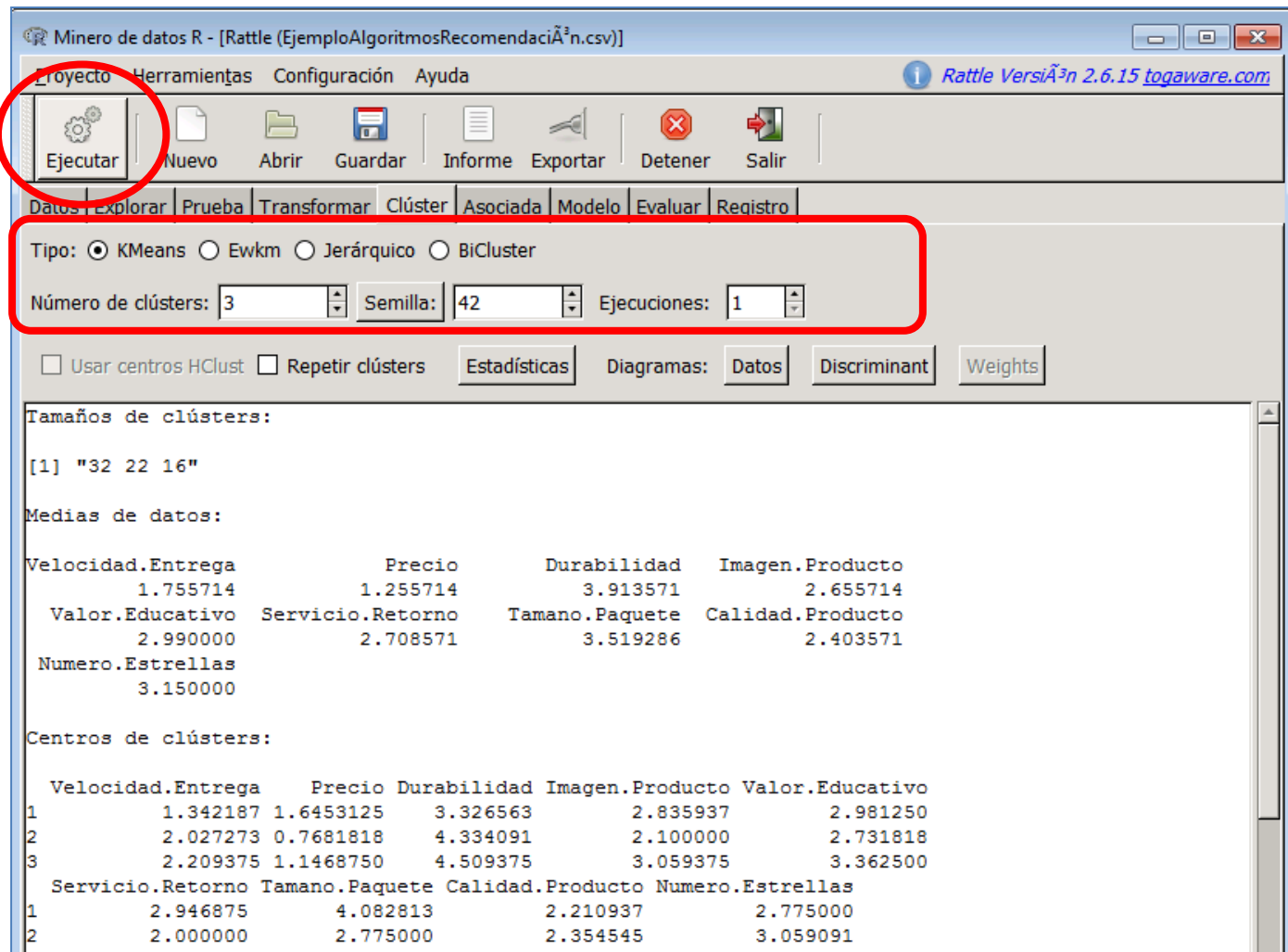
☒ Partición 70/15/15 Semilla: 42 Ver Editar

☒ Entrada ☒ Ignorar Calculadora de peso:

Tipo de datos de destino: ☒ Automática ☐ Categórica ☐ Numérica ☐ Supervivencia

No.	Variable	Tipo de datos	Entrada	Destino	Riesgo	Ident	Ignorar	Weight	Comentario
1	X	Categórica	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Única: 100
2	Velocidad.Entrega	Numérica	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Única: 47
3	Precio	Numérica	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Única: 45
4	Durabilidad	Numérica	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Única: 47
5	Imagen.Producto	Numérica	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Única: 40
6	Valor.Educativo	Numérica	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Única: 32
7	Servicio.Retorno	Numérica	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Única: 29
8	Tamaño.Paquete	Numérica	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Única: 46
9	Calidad.Producto	Numérica	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Única: 30
10	Numero.Estrellas	Numérica	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Única: 35

# Generando el modelo k-means



The screenshot shows the Rattle GUI for generating a KMeans model. The 'Ejecutar' button is highlighted with a red circle. The 'Tipo' section shows 'KMeans' selected. The 'Número de clústers' is set to 3, 'Semilla' is 42, and 'Ejecuciones' is 1. The 'Estadísticas' tab is active, displaying the following results:

Tamaños de clústers:

```
[1] "32 22 16"
```

Medias de datos:

Velocidad.Entrega	Precio	Durabilidad	Imagen.Producto
1.755714	1.255714	3.913571	2.655714
Valor.Educativo	Servicio.Retorno	Tamano.Paquete	Calidad.Producto
2.990000	2.708571	3.519286	2.403571
Numero.Estrellas	3.150000		

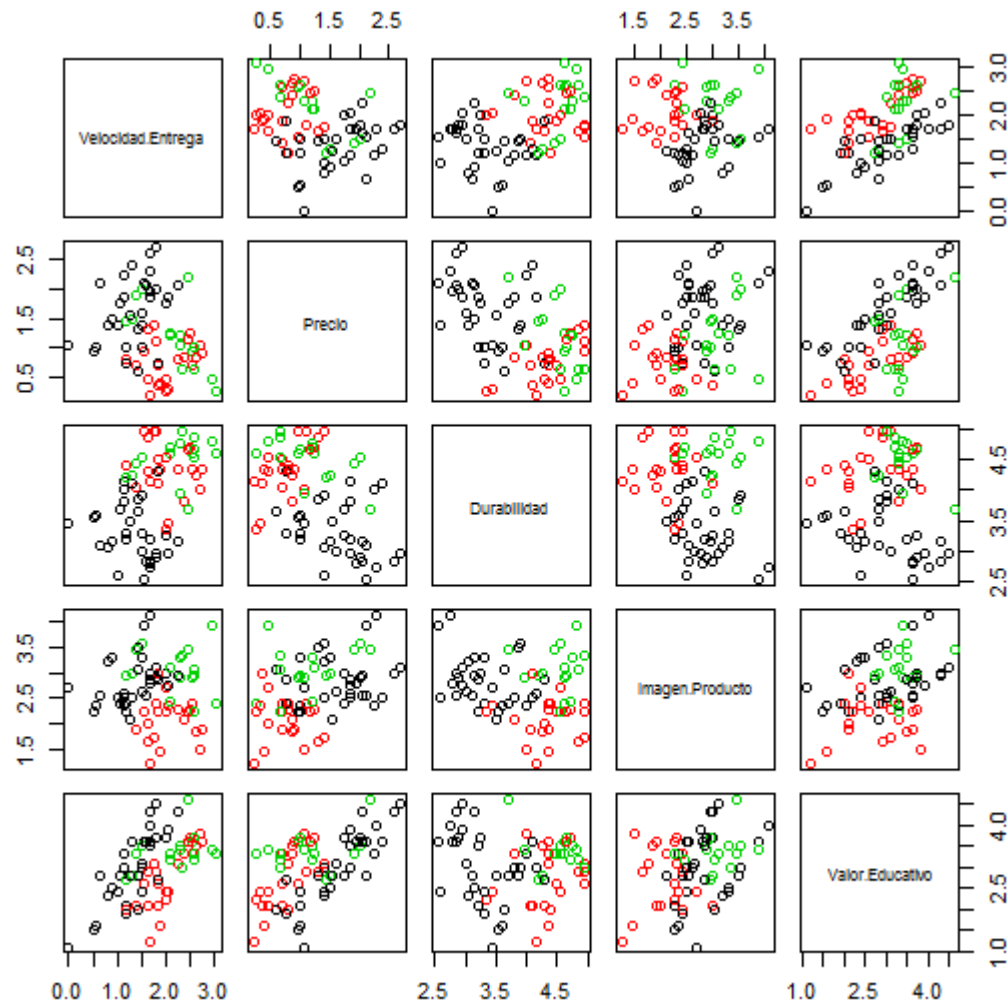
Centros de clústers:

	Velocidad.Entrega	Precio	Durabilidad	Imagen.Producto	Valor.Educativo
1	1.342187	1.6453125	3.326563	2.835937	2.981250
2	2.027273	0.7681818	4.334091	2.100000	2.731818
3	2.209375	1.1468750	4.509375	3.059375	3.362500

	Servicio.Retorno	Tamano.Paquete	Calidad.Producto	Numero.Estrellas
1	2.946875	4.082813	2.210937	2.775000
2	2.000000	2.775000	2.354545	3.059091



# Opción “Datos”



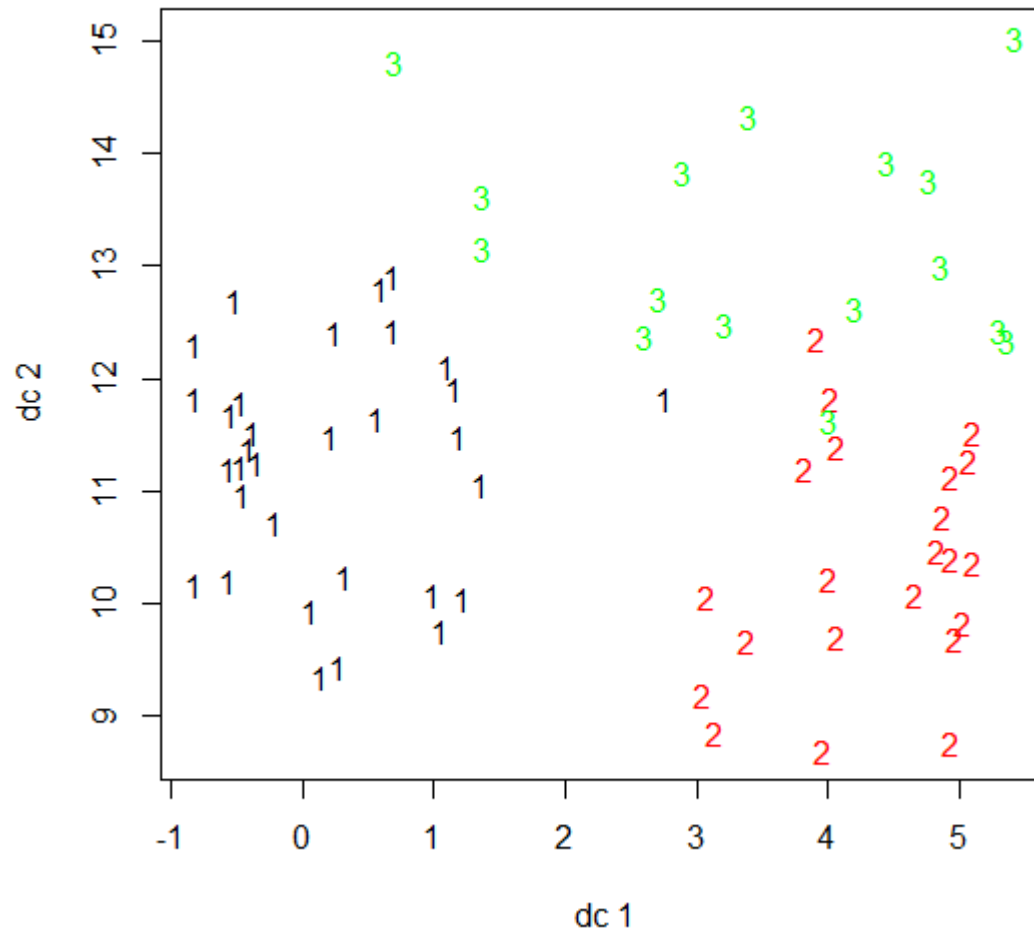
Rattle 2012-ago-17 14:32:06 Oldemar



**oldemar rodríguez**  
CONSULTOR en MINERÍA DE DATOS

# Opción “Discriminat”

Discriminant Coordinates Ejemplo Algoritmos Recomendación



*Gracias....*



**oldemar** **rodríguez**

CONSULTOR en M1N&R14 D& D4T0S