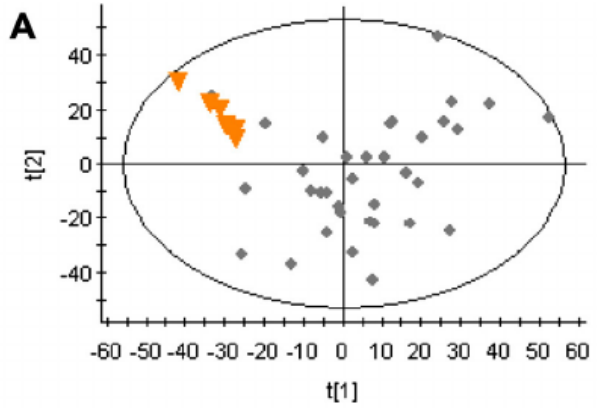
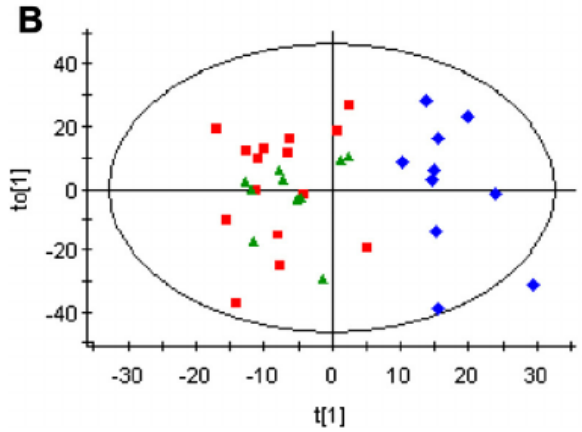
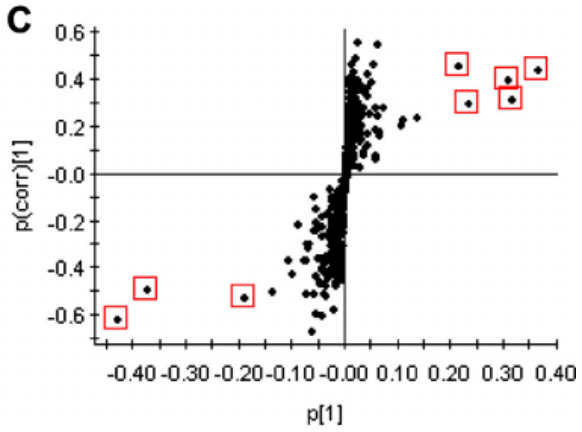
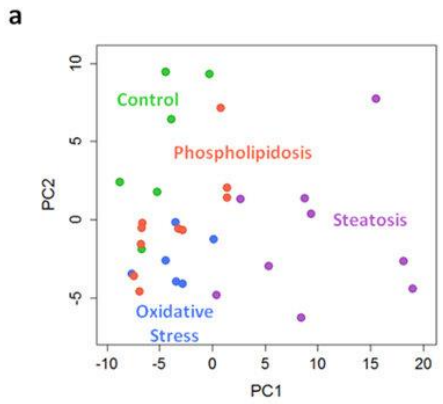
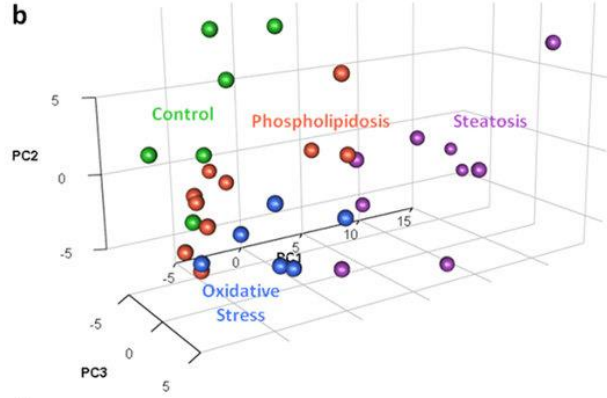
Introducción al Análisis Multivariados

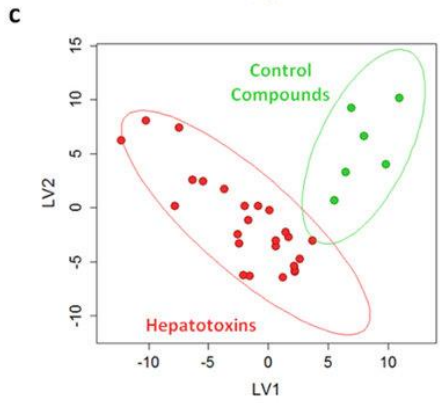


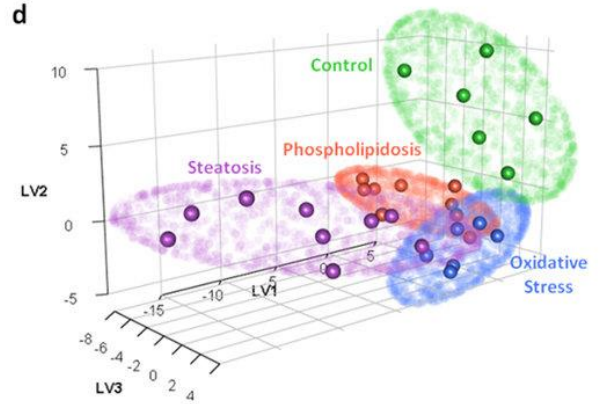












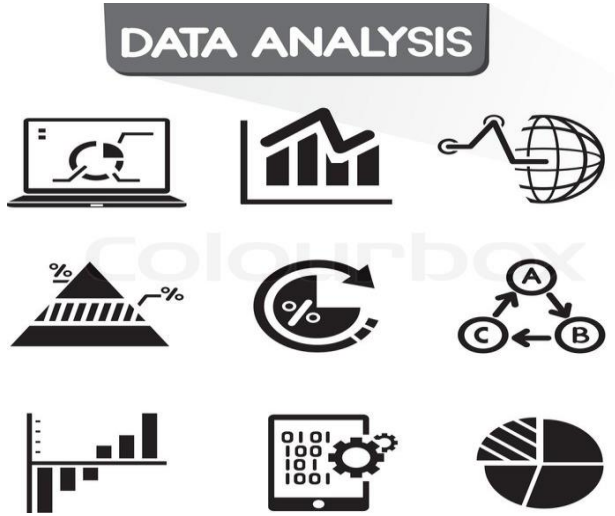
Oscar Centeno Mora

Análisis en una variable

- ¿Cuáles son los principales análisis que realizamos en una variable X1 (cuantitativa y cualitativa)?

Análisis en una variable

- Números relativos (razón, proporción, tasas, indicadores, etc.)
- Medidas de posición (promedio, percentiles, moda, mediana, etc.)
- Medidas de variabilidad (rango, percentil intercuartil, desviación media, desviación estándar, etc.)
- Probabilidades.
- Medidas de estimación
- Prueba de hipótesis.
- Medidas de visualización



Análisis multivariados o en varias variables

- ¿Cuáles son los análisis que conocemos para varias 2 o más variables X1,X2,…, Xn (cuantitativas y cualitativas)?

Análisis multivariados o en varias variables

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots \beta_n X_n + \varepsilon$$

$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots \beta_n X_n + \varepsilon$

- y: Dependent Variable
- β_0, β_1, β_2: Coefficients
- X_1, X_2: Explanatory Variables
- ε: Random Error Term/Residuals

Example: - Suppose you want to both model and predict residential burglary (RES_BURG) for the census tracts in your community. You've identified median income (MED_INC), the number of vandalism incidents (VAND) and the number of household units (HH_UNITS) to be key explanatory variables. The regression equation would have the elements below.

$$RES\_BURG = \beta_0 + \beta_1 * (MED\_INC) + \beta_2 * (VAND) + \beta_3 * (HH\_UNITS) + \varepsilon$$

Análisis multivariados o en varias variables

- Análisis de covariancia, correlación, etc.
- Análisis por regresión (paramétrica, no paramétrica, semi paramétrica, etc.).
- Análisis experimental (comparación de medias, ANOVA, GLM, etc.)
- Análisis por descomposiciones.
- Análisis por discriminación.
- Análisis por agrupamiento.
- Análisis por clasificación y predicción.
- Etc….

Análisis multivariados descriptivos y supervisados

Análisis multivariados

Análisis descriptivas (búsqueda de patrones)

Análisis predictivos (predicciones en los casos --- Minería de datos)

Principios del análisis de datos multivariado

Principios multivariados: datos

- Conjuntos de unidades y variables por analizar: matriz de valores.

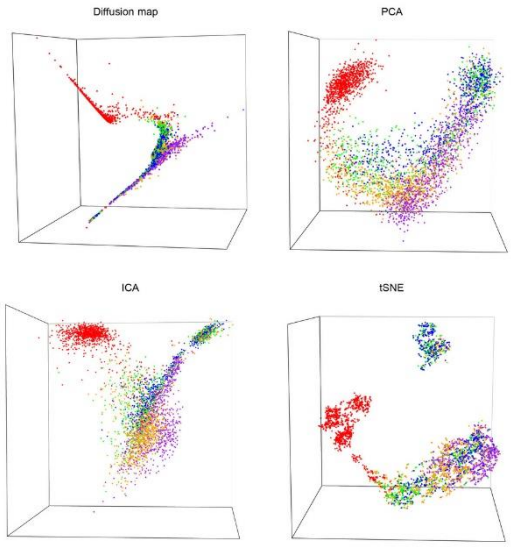
- Tipo de eventos es común en todos los campos: psicología, ingeniería, educación, física, química, economía, etc.

- Representación en un archivo de datos:

Unit	Variable 1	...	Variable q
1	x_{11}	...	x_{1q}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	x_{n1}	...	x_{nq}

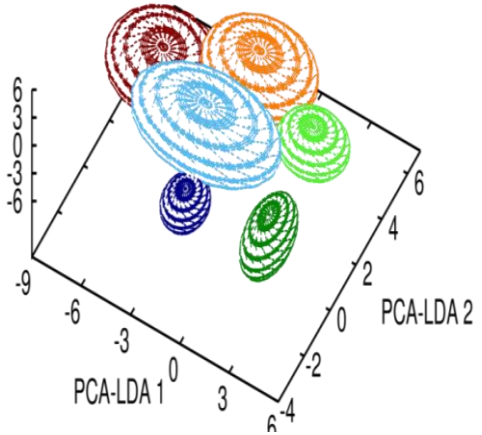
Principios multivariados: análisis

- Análisis multivariados descriptivo: análisis que relacionan variables independientes (*relación simétrica entre variables*), para buscar algún tipo de patrón común.



- No hay interés en hacer inferencia estadística.

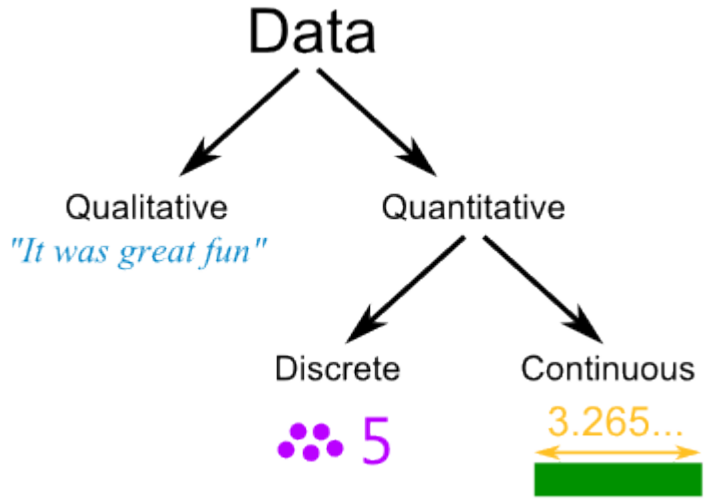
- Los análisis multivariados se exponen tanto de forma tabular y gráfica, siendo la última la más utilizada.



- El análisis multivariado exploratorio da prioridad a la información recolectada, y no al proceso inferencial. Es un proceso meramente analítico para y por los datos.

Tipos de variables y problemas asociados

- Los análisis multivariados toman tanto variables cualitativas como cuantitativas.



- La técnica se adecua según el objetivo y tipo de variable.

- Los valores faltantes provocan la eliminación de una unidad en el archivo de datos.

- Los valores faltantes provocan perdida de información y problemas de cálculos matriciales.

Covariancia, correlación y distancia

- El análisis multivariado implica relaciones entre las variables o la posible cercanía que pueden haber entre estas. En ciertos casos pueden ser los dos.

- El caso que interesa en el análisis multivariado es como calificamos la relación entre las variables y como podemos medir las distancias entre variables con diferentes unidades medida.

- Para esto debemos recurrir a la covariancia, correlación y el análisis de distancias.

Covariancia

- ¿Qué es una variancia? Y, ¿qué es una covariancia?
- ¿Cuál es la covariancia de una variable con ella misma?
- La covariancia en dos variables aleatorias sirve para medir el grado de dependencia linear. Su notación matemática es la siguiente:

$$\mathrm{Cov}(X_i, X_j) = \mathrm{E}(X_i - \mu_i)(X_j - \mu_j)$$

- En una matriz de variables, la covariancia se expresa como sigue:

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1q} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{q1} & \sigma_{q2} & \dots & \sigma_q^2 \end{pmatrix}$$

σ_i^2 Variancia

σ_{ij} Covariancia

Correlación

- La dificultad para interpretar la covariancia se facilita cuando se utiliza la correlación.
- Esta es la división de una covariancia respeto a las respectivas desviaciones estándares. Su notación matemática es:

$$\rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$$

- En una matriz de datos, se utiliza una matriz de diagonalización D y la matriz de correlación S, para obtener la matriz de correlación R.

$$\mathbf{D}^{-1/2} = \operatorname{diag}(1/s_1, \ldots, 1/s_q)$$

$$\mathbf{S} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \ldots & \sigma_{1q} \\ \sigma_{21} & \sigma_2^2 & \ldots & \sigma_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{q1} & \sigma_{q2} & \ldots & \sigma_q^2 \end{pmatrix}$$

$$\mathbf{R} = \mathbf{D}^{-1/2}\mathbf{S}\mathbf{D}^{-1/2}$$

Distancia

- Los análisis de escalamiento multidimensional y análisis por agrupamiento se fundamentan en las distancias de las unidades de los datos.

- Para dos sujetos *i* e *j*, para una determinada variable se toman las distancias que existen para ese par de individuos.

- Existen diversos tipos de distancias (Mahalanobis, Hellinger, Manhattan, etc.), pero la más común es la distancia Euclidiana:

$$d_{ij} = \sqrt{\sum_{k=1}^{q} (x_{ik} - x_{jk})^2}$$

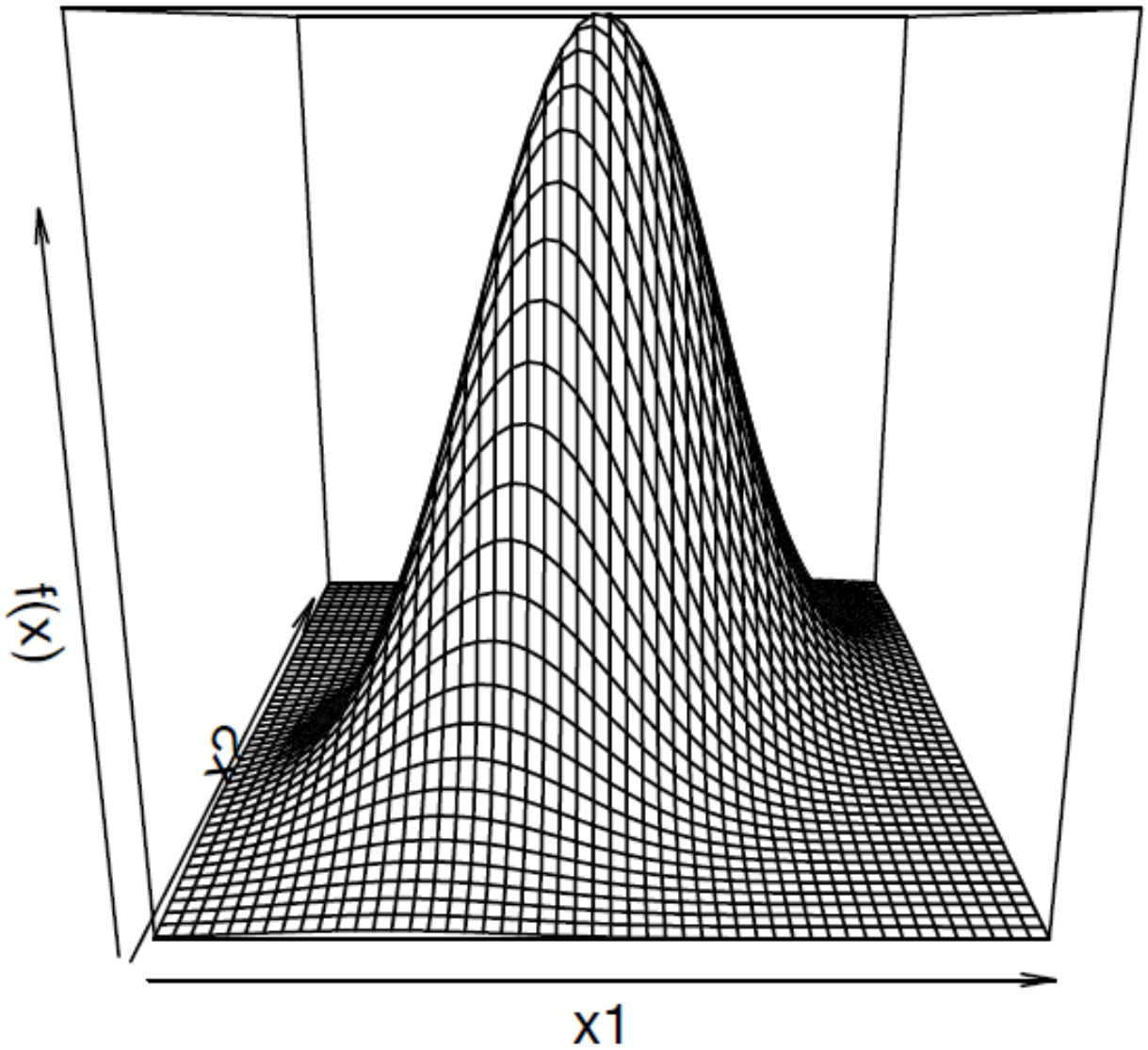
Función normal multivariable

- La función normal multivariada juega un rol importante en algunos análisis multivariados.
- Muchos supuestos implícitos consideran que los datos provienen de distribuciones normales.
- Una función normal multivariada de dos variables se expresa como:

$$f(\mathrm{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = (2\pi)^{-q/2}\det(\boldsymbol{\Sigma})^{-1/2}\exp\left\{-\frac{1}{2}(\mathrm{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\mathrm{x}-\boldsymbol{\mu})\right\}$$

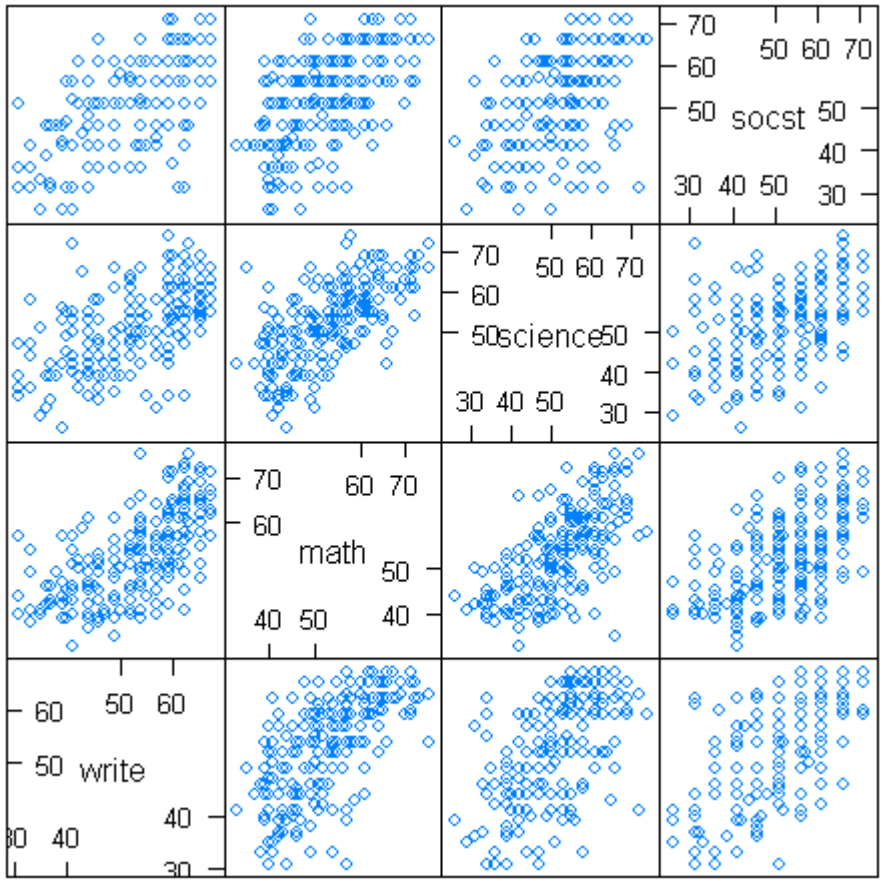
- Una función normal multivariada se presenta como sigue en la siguiente diapositiva.

Función normal multivariable

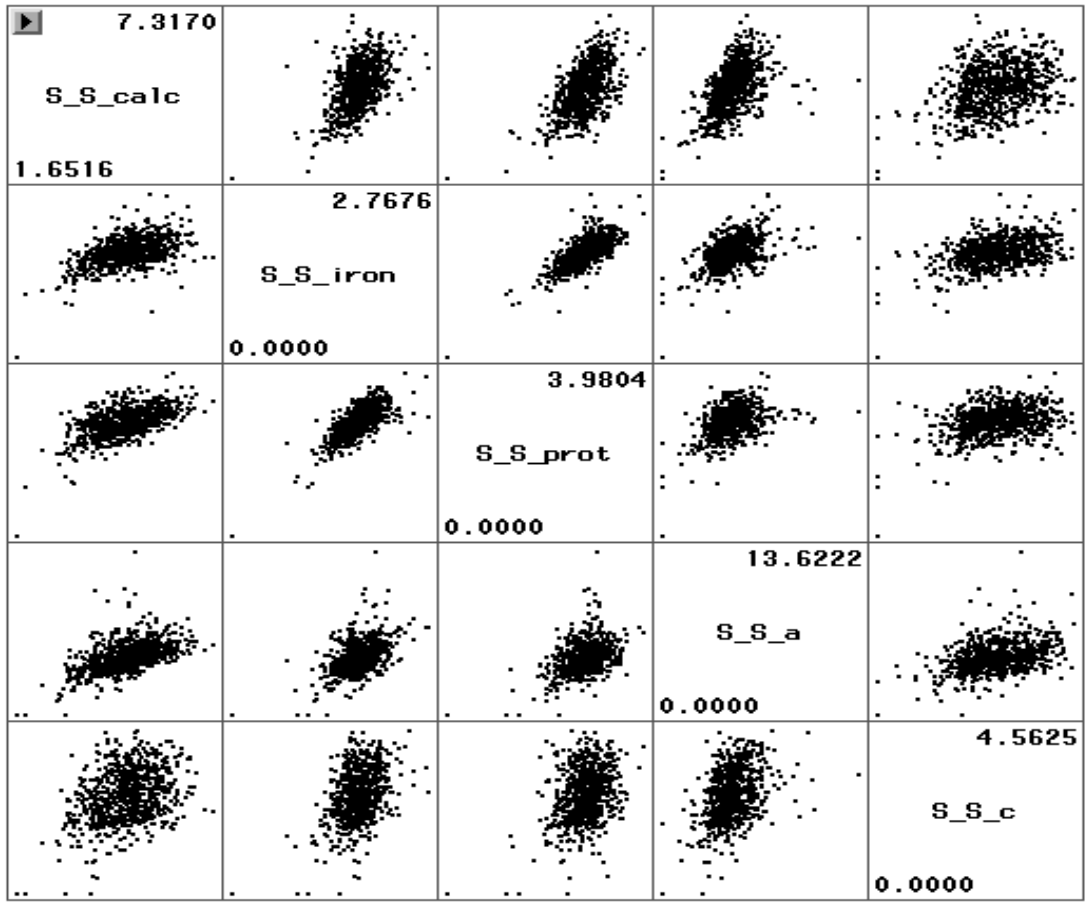


Visualización de datos multivariados

- Principales gráficos multivariados: contornos, matrices de gráficos, correlogramas, gráficos bi-tridimensionales.

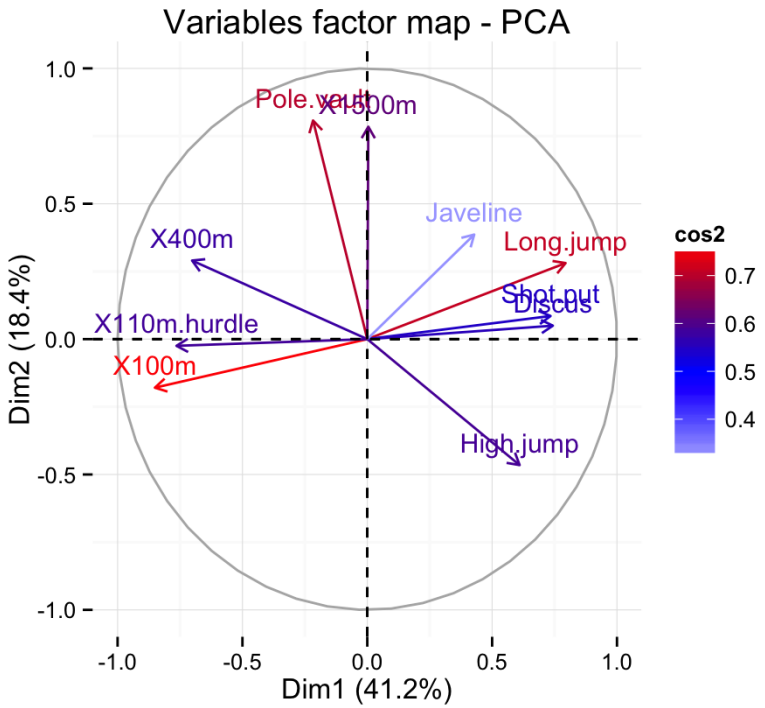


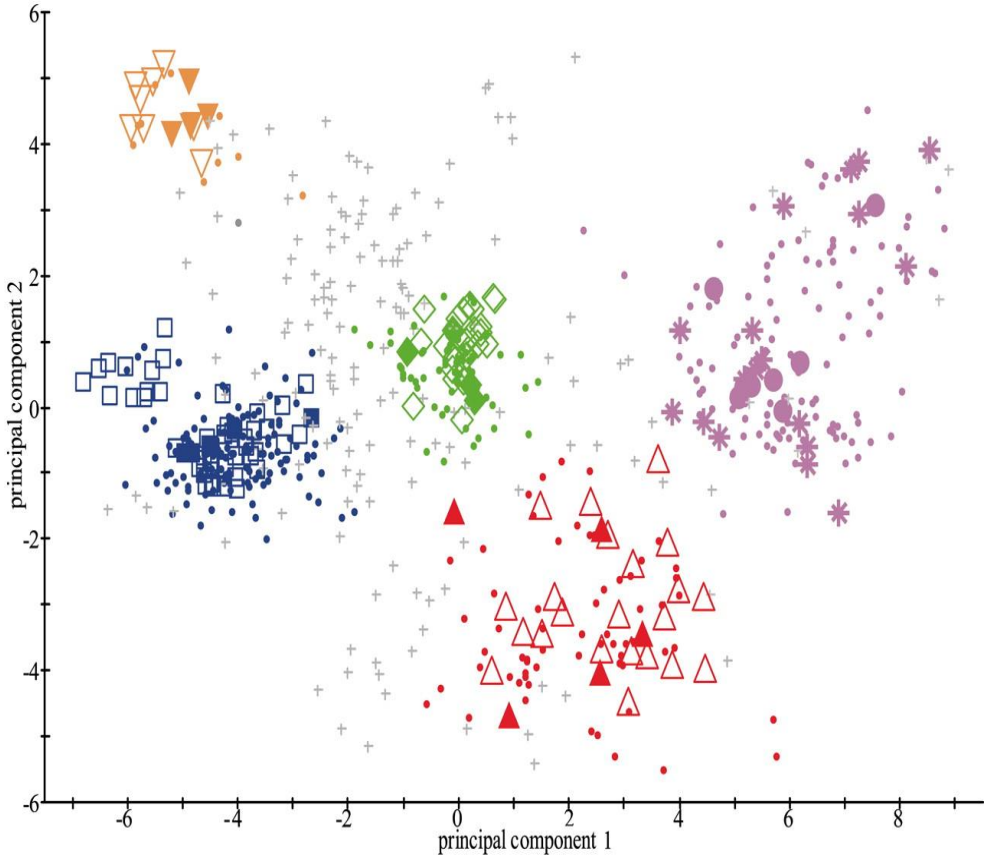
Scatter Plot Matrix



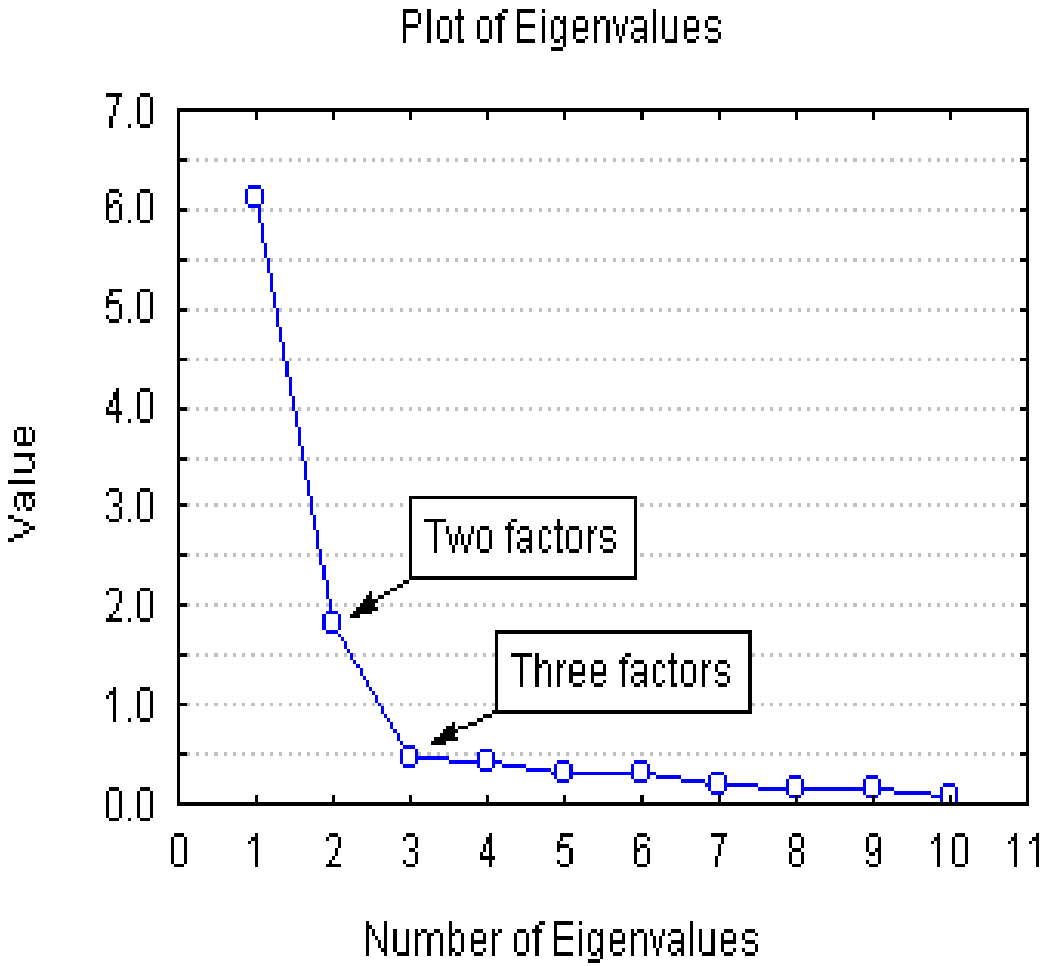
Análisis por componentes principales

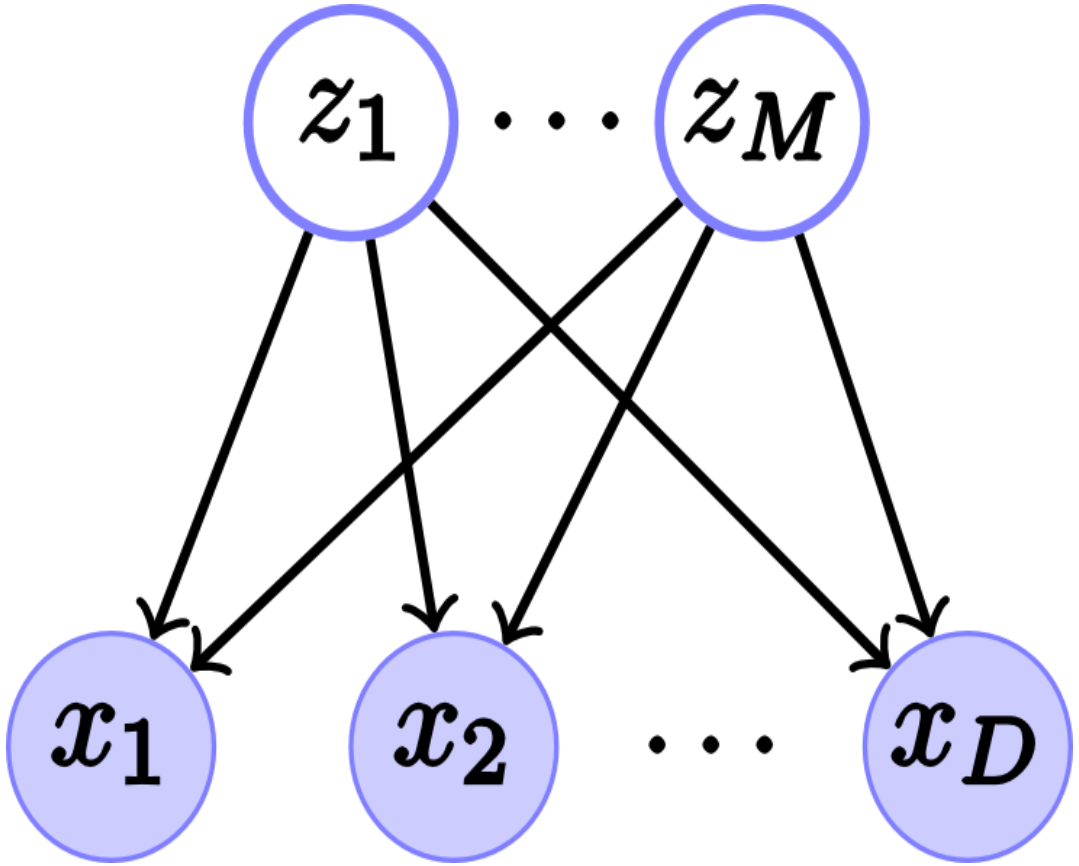
- Reducción de variables y representación de sujetos, variables y ambos según los componentes.





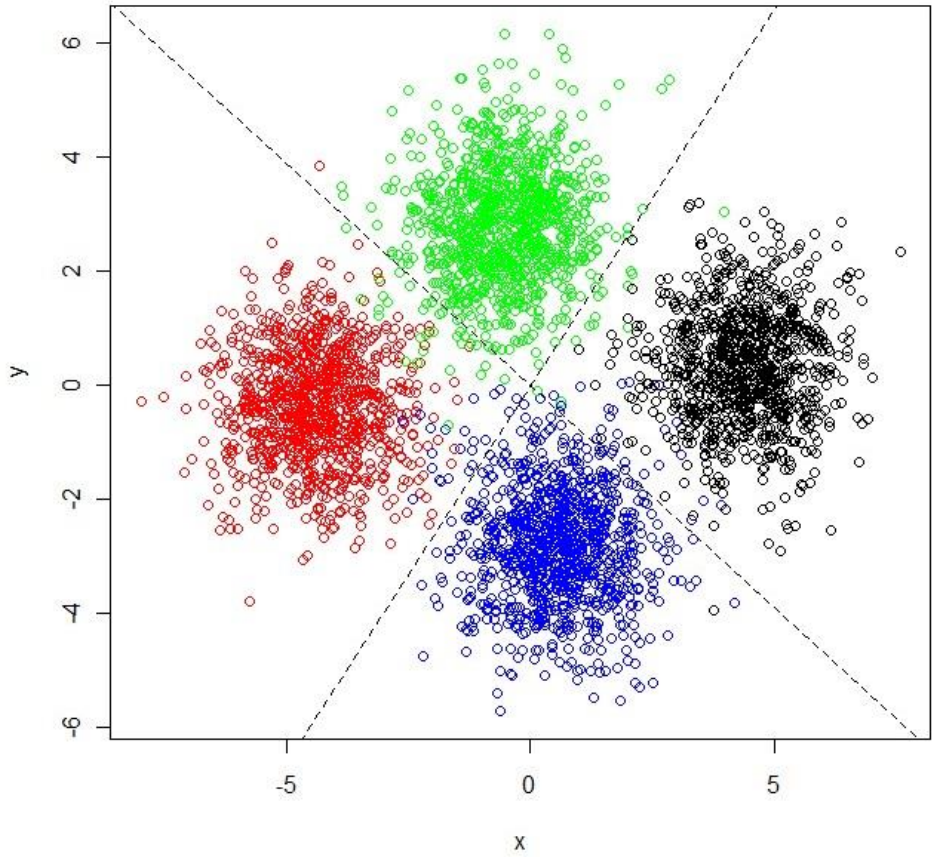
Análisis factorial

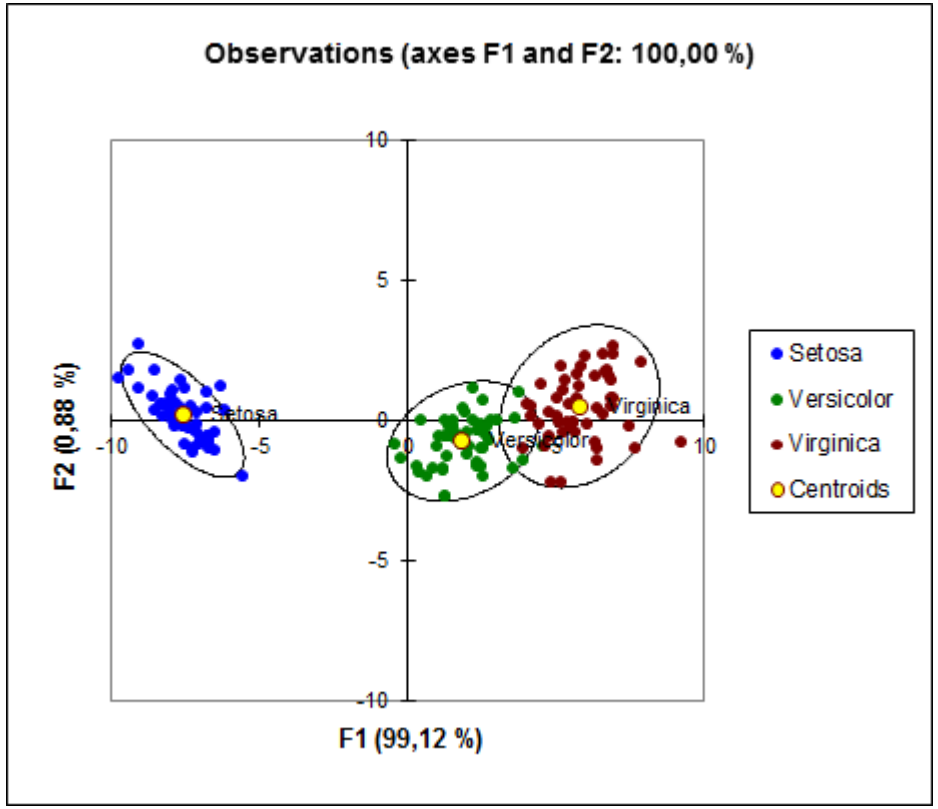




Análisis discriminante

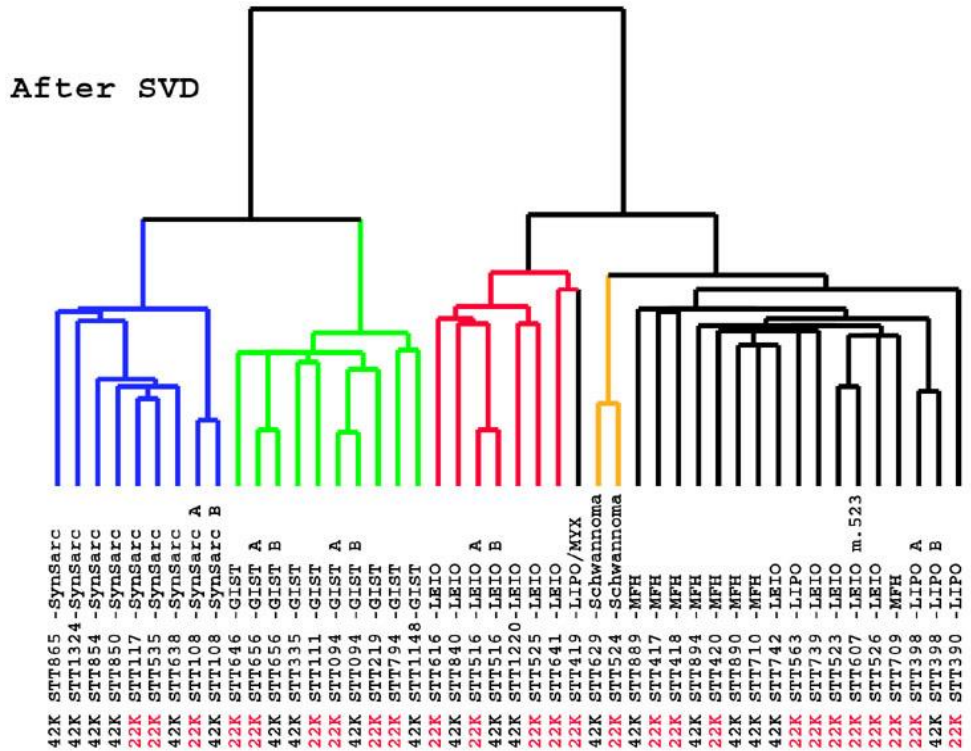
- Busca separar a discriminar a las unidades según las características de estos.

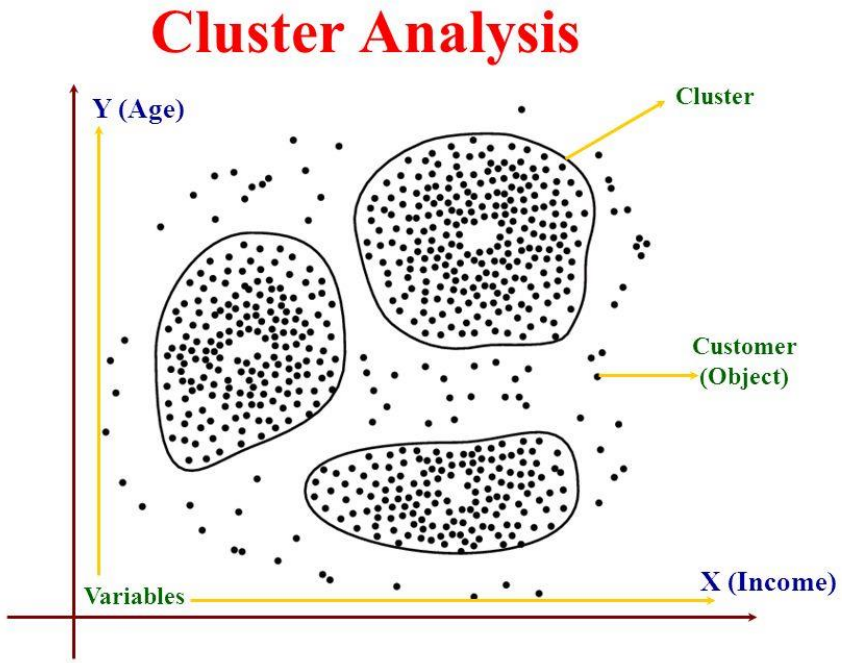




Análisis por agrupamiento

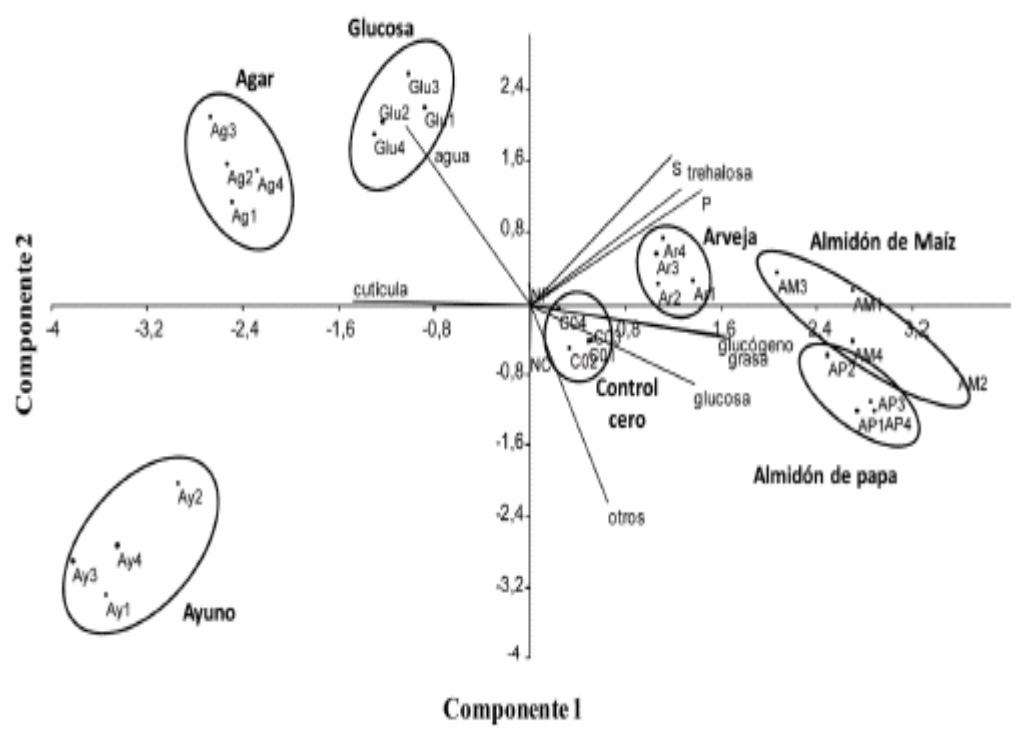
- A partir de las observaciones en las *k* variables de los casos, busca agrupar o juntar a los individuos.

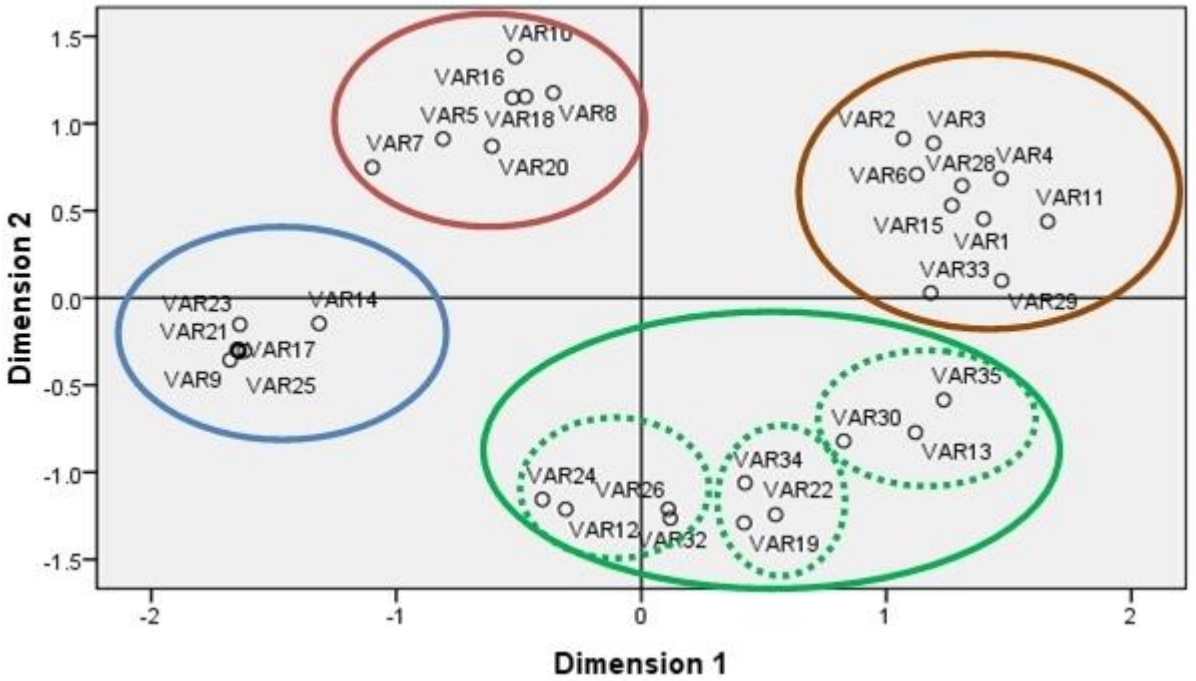




Escalamiento multidimensional

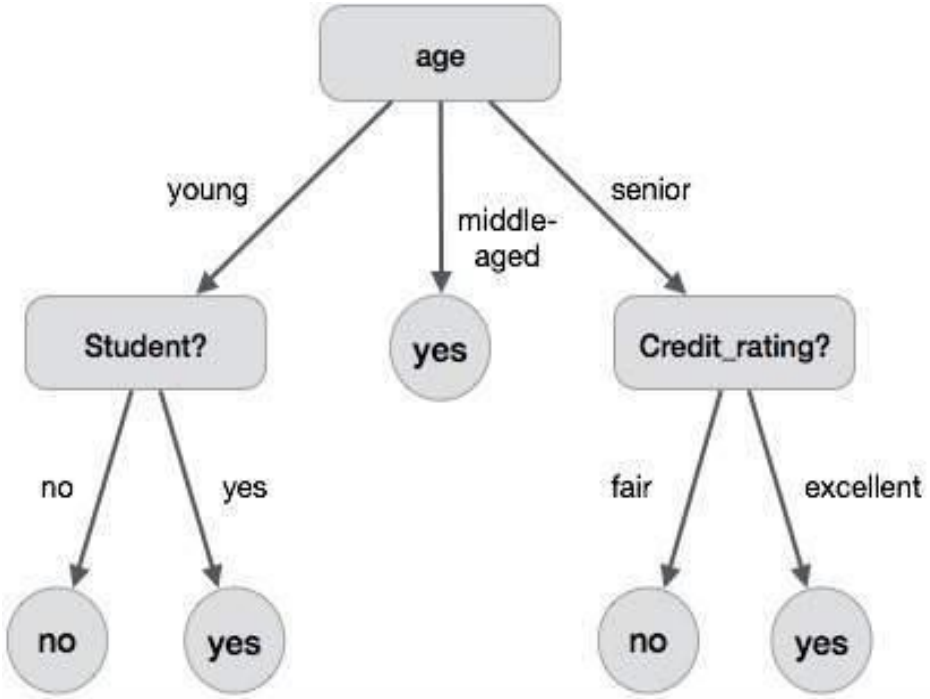
- Se trata de un procedimiento para dibujar mapas sobre los que representan geométricamente, en forma de puntos, un conjunto de objetos. Este análisis utiliza variables cualitativas.

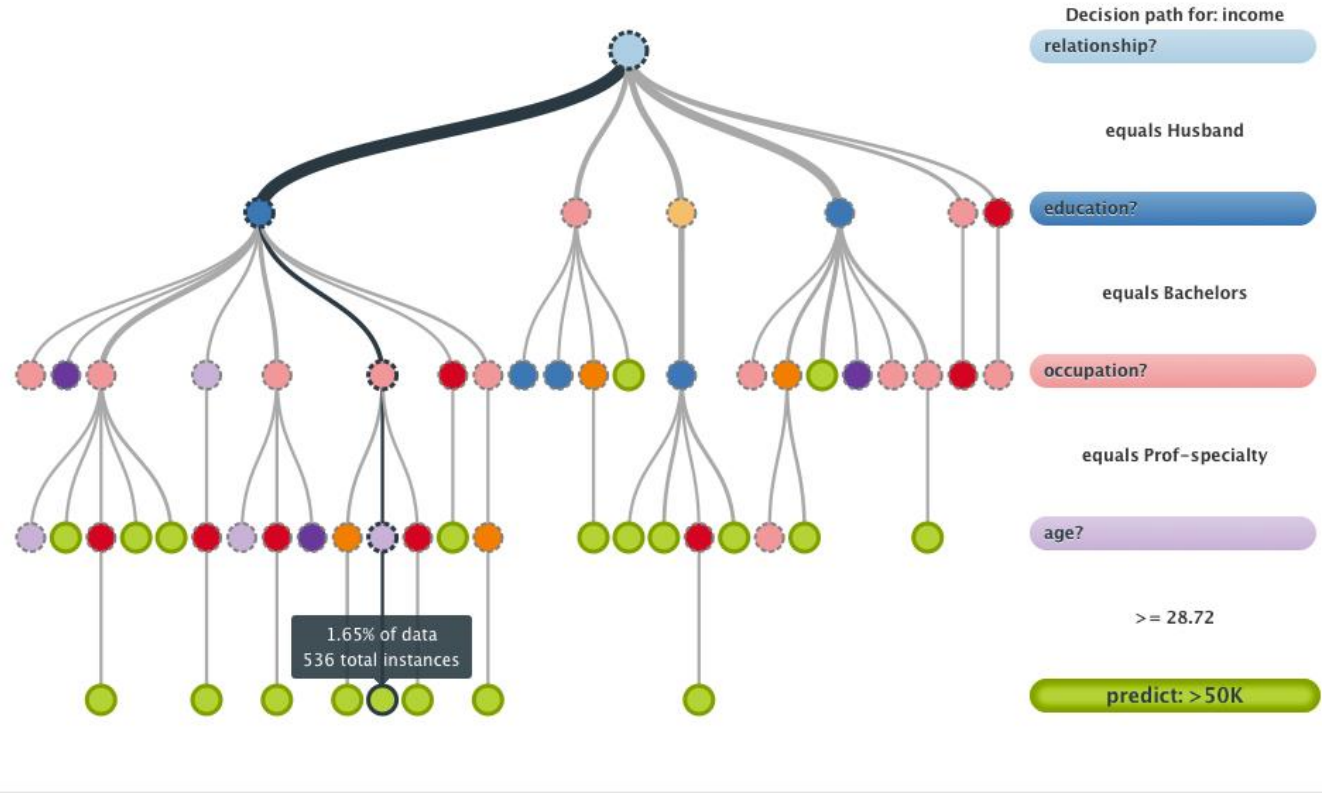




Árboles de decisión

- Para clasificar y predecir, se utiliza para predecir cierta característica Y, para un conjunto de variables X1,…,Xp.





¿Nivel de programación en R?