# Análisis multivariado y reducción de dimensiones

**Hugo Andrés Dorado**

Científico de datos

hugo.doradob@gmail.com

# Análisis multivariado

- Analizar simultáneamente un conjunto de datos en el que se han medido varias variables.
- Se desea obtener un mejor entendimiento del fenómeno, traspasando las limitaciones de los métodos bivariados y univariados.
- Ayudar al analista a tomar decisiones.

# Clasificación según los métodos

- Métodos de dependencia.
  - Regresión lineal multiple
  - Análisis de varianza.
- Métodos de independencia.
  - Análisis de componentes principales
  - Análisis de correspondencia multiple
- Métodos de estructurales.
  - Análisis factorial multiple/

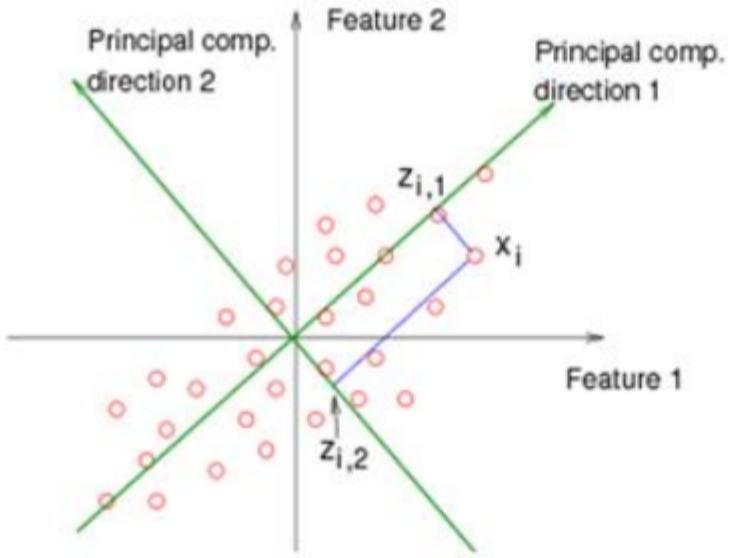
# Análisis de componentes principales

- Describir un conjunto de datos en términos de nuevas variables, componentes, que no están correlacionadas.
- Una muestra con n individuos cada uno con p variables (X1, X2, …, Xp) de p dimensiones. PCA permite encontrar un número de factores subyacentes (z)

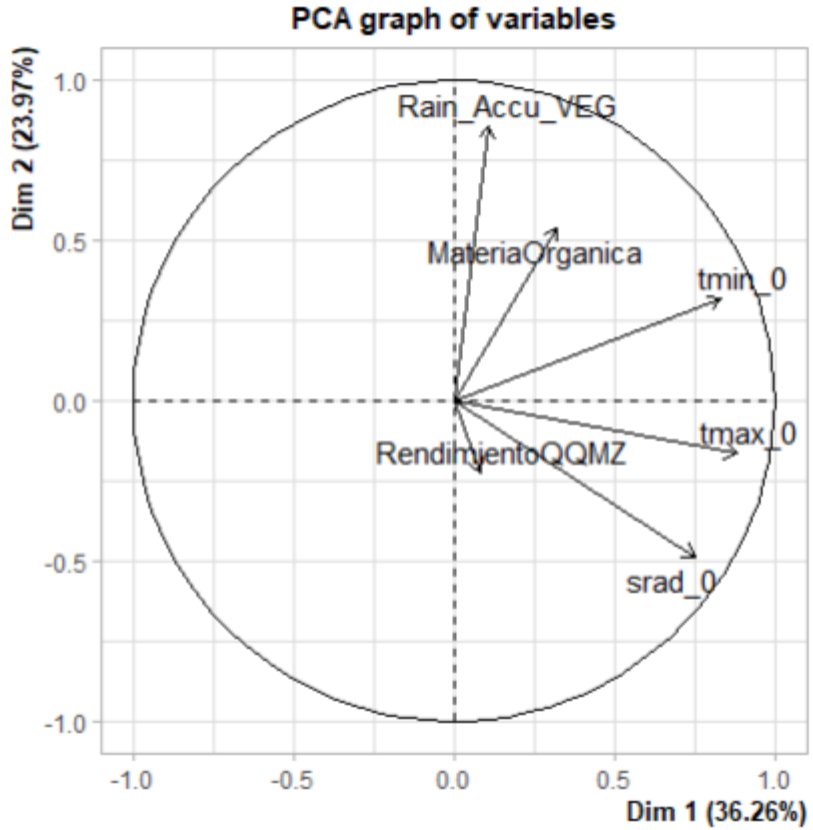
$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}_{nxp} \rightarrow \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1z} \\ w_{21} & w_{22} & \cdots & w_{2z} \\ \vdots & \vdots & & \vdots \\ w_{n1} & w_{n2} & \cdots & w_{nz} \end{bmatrix}_{nxz}$$

$$p < z$$

# Análisis de componentes principales



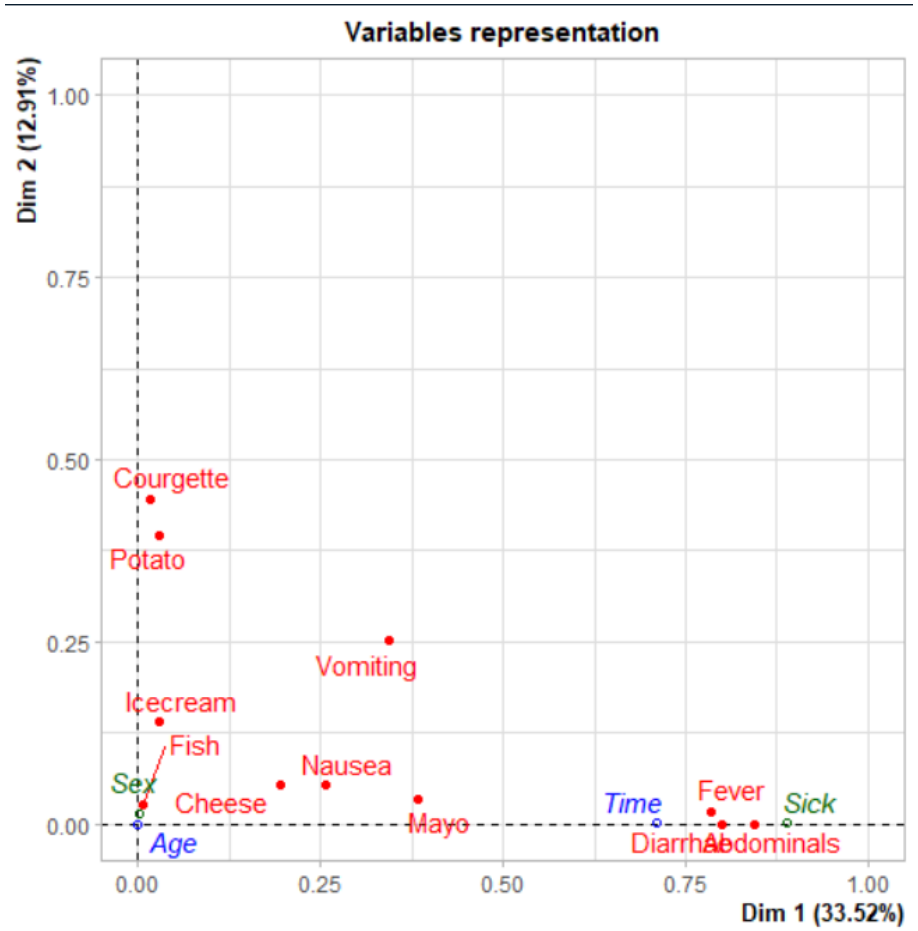

Fuente: https://online.stat.psu.edu/stat857/node/37/



# Análisis de correspondencia multiple

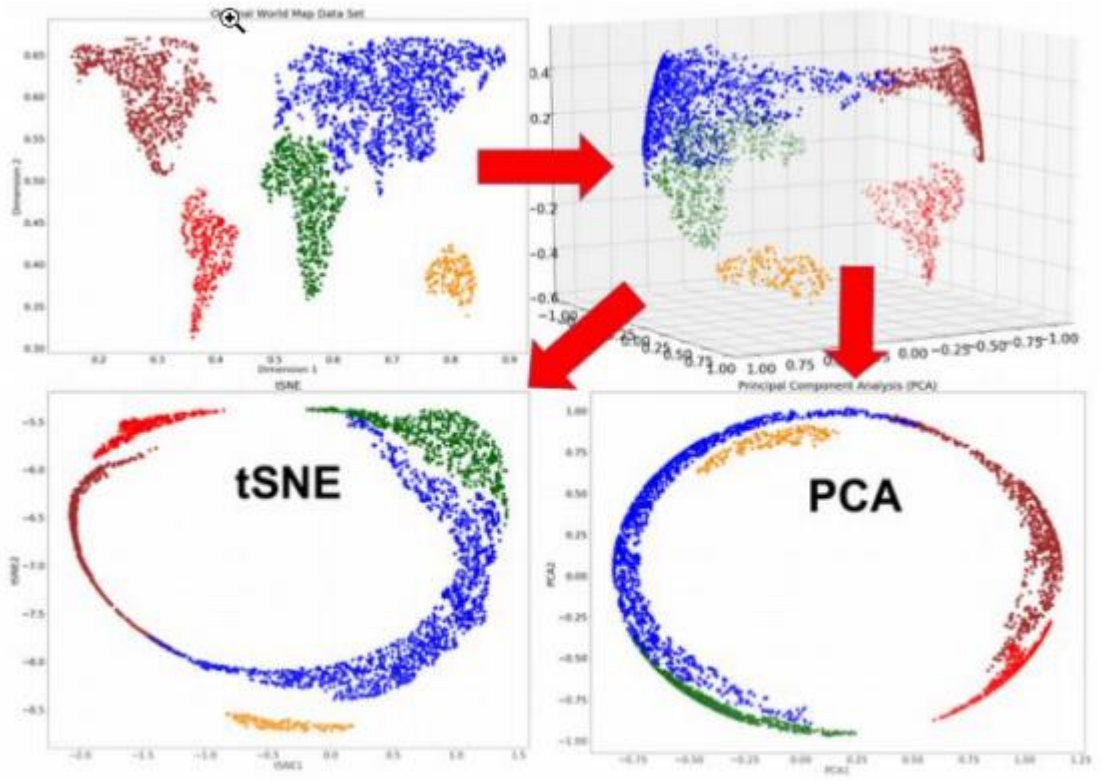
```
##     Nausea Vomiting Abdominals   Fever   Diarrhae   Potato
## 1 Nausea_y  Vomit_n     Abdo_y Fever_y Diarrhea_y Potato_y
## 2 Nausea_n  Vomit_n     Abdo_n Fever_n Diarrhea_n Potato_y
## 3 Nausea_n  Vomit_y     Abdo_y Fever_y Diarrhea_y Potato_y
```

```
'data.frame':   55 obs. of  15 variables:
 $ Age        : int  9 5 6 9 7 72 5 10 5 11 ...
 $ Time       : int  22 0 16 0 14 9 16 8 20 12 ...
 $ Sick       : Factor w/ 2 levels "Sick_n","Sick_y": 2 1 2 1
 $ Sex        : Factor w/ 2 levels "F","M": 1 1 1 1 2 2 1 1 2
 $ Nausea     : Factor w/ 2 levels "Nausea_n","Nausea_y": 2 1
 $ Vomiting   : Factor w/ 2 levels "Vomit_n","Vomit_y": 1 1 2
 $ Abdominals: Factor w/ 2 levels "Abdo_n","Abdo_y": 2 1 2 1
 $ Fever      : Factor w/ 2 levels "Fever_n","Fever_y": 2 1 2
 $ Diarrhae   : Factor w/ 2 levels "Diarrhea_n","Diarrhea_y":
 $ Potato     : Factor w/ 2 levels "Potato_n","Potato_y": 2 2
 $ Fish       : Factor w/ 2 levels "Fish_n","Fish_y": 2 2 2 2
 $ Mayo       : Factor w/ 2 levels "Mayo_n","Mayo_y": 2 2 2 1
 $ Courgette  : Factor w/ 2 levels "Courg_n","Courg_y": 2 2 2
 $ Cheese     : Factor w/ 2 levels "Cheese_n","Cheese_y": 2 1
 $ Icecream   : Factor w/ 2 levels "Icecream_n","Icecream_y":
```



# T-SNE – (T-distributed Stochastic Neighbor Embedding)

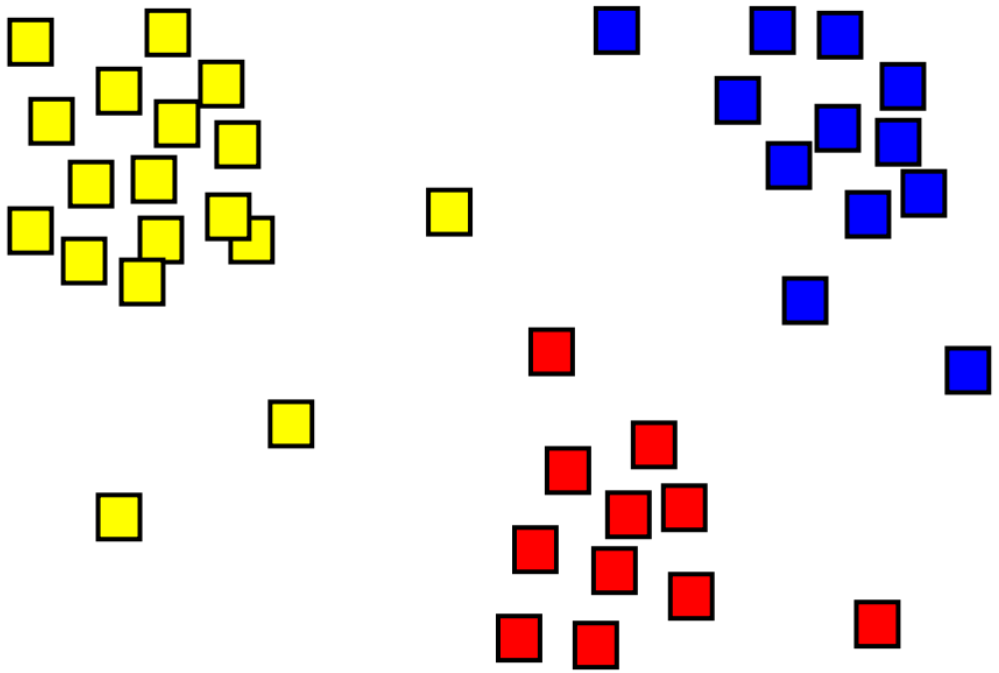
- Es un método para la reducción de dimensionalidad y visualización basado en métodos no lineales.
- Se construye a partir de asignación de distribuciones de probabilidades en dimensiones altas de tal manera que los objetos (datos) con distribución similar tienen una probabilidad alta.




[Oskolkov]

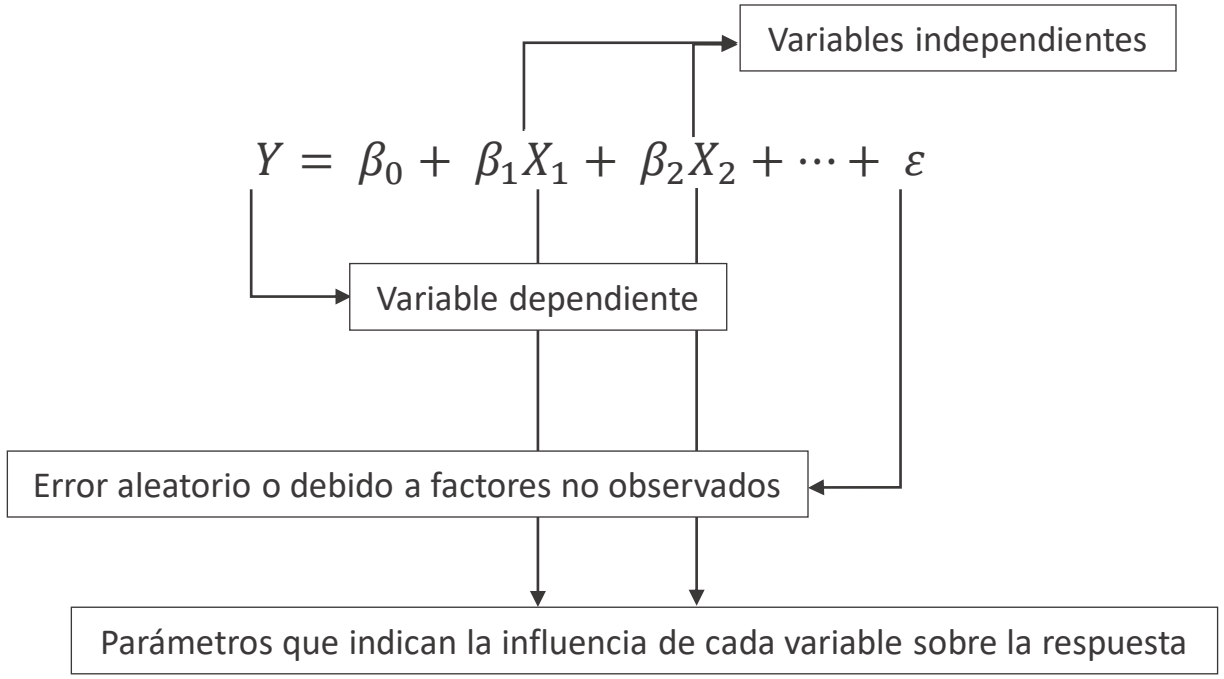
La **perplejidad** está relacionada con el número de vecinos más cercanos que se utilizan en otros algoritmos de aprendizaje de colector. Normalmente, los conjuntos de datos más grandes necesitaban una **perplejidad** mayor.

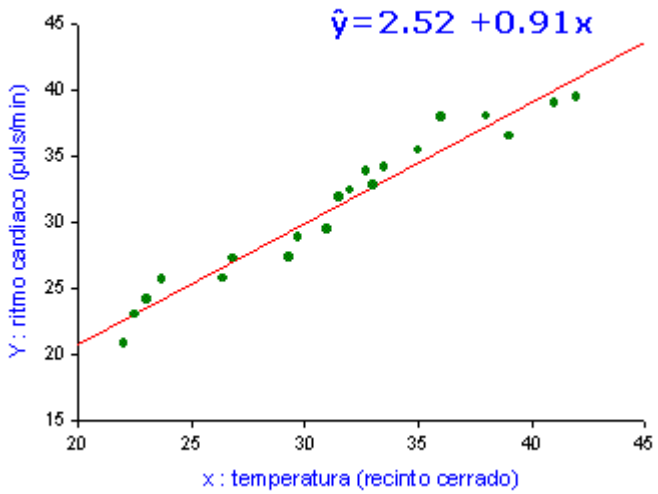
# Análisis cluster – Análisis de conglomerados



# Regresión lineal multiple

Es un modelo matemático que busca ajustar una ecuación lineal que maximize las relaciones entre una variable dependiente ‘Y’ y un conjunto de variables independientes (X1,X2,…,Xn) y un término de error.




Fuente: e-stadistica.bio.ucm.es

**Principales supuestos**

- Relaciones lineales entre variables
- Las mediciones deben ser independientes
- Los errores deben tener varianza constante
- Los errores deben seguir una distribución normal

# Análisis de varianza

El objetivo principal de muchos experimentos consiste en determinar el efecto que sobre alguna variable dependiente **Y** tienen distintos niveles de algún factor **X** (variable independiente y discreta).

El efecto se evalúa a través de la comparación de las medias de cada nivel de la variable discreta X.

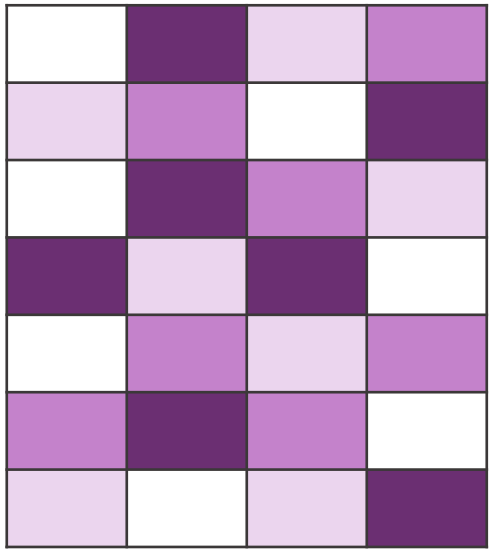
Fuente: https://www.uoc.edu/in3/emath/docs/ANOVA.pdf

$$Y = \mu + \tau + \varepsilon$$

Hay algún efecto sobre el rendimiento (Y) de acuerdo a la cantidad de nitrógeno (X) aplicada en mi finca?, donde se presentan las diferencias?

Cantidad aplicada de nitrógeno

- 0 kg
- 5 kg
- 10 kg
- 15 kg



# Prácticas en R

# Gracias!

**Hugo Andrés Dorado.**

Científico de datos

hugo.doradob@gmail.com

Conocimiento generado a partir de proyectos de:

Alianza

Bioversity International

CIAT
Centro Internacional de Agricultura Tropical
Desde 1967 Ciencia para cultivar el cambio

CLIMATE CHANGE AGRICULTURE AND FOOD SECURITY