# Data Mining

## Análisis de valores atípicos

# Análisis de Outliers

"Un *outlier* es una observación que se desvía tanto de las otras observaciones como para despertar sospechas sobre que fue generado por un mecanismo diferente"

*D. Hawkins. Identification of Outliers*

- Análisis de outliers.
  - Tipos de outliers
  - Revisión de métodos univariados: IQR, Boxplot, z-score….
  - Métodos multivariados: Local Outlier Factor (LOF)

# Tipos de outlier

**Univariado**

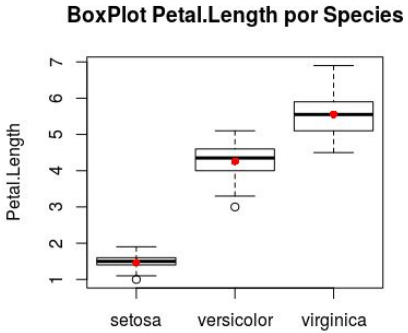
- Son valores **atípicos** que podemos encontrar en una simple variable.
- El problema de los enfoques univariados es que son buenos para detección de extremos pero no en otros casos.

**Multivariado**

- Los valores atípicos multivariados se pueden encontrar en un espacio n-dimensional.
- Para detectar valores atípicos en espacios n-dimensionales es necesario ajustar un modelo.

# Outliers

## Métodos Univariados




- IQR: Analizar los valores que están por fuera del IRQ
- Z-score.
- Identificar valores extremos a partir de 1, 2 o 3 desvíos de la media.

## Métodos Multivariados

- Análisis globales: Clustering.
  - Utilizando medidas de distancia como Mahalanobis.
- Local Outlier Factor (LOF)
  - Es un método de detección de outliers basado en distancias.
  - Calcula un score de *outlier* a partir de una distancia que se normaliza por densidad.
- Métodos basados en ensambles: IsolationForest

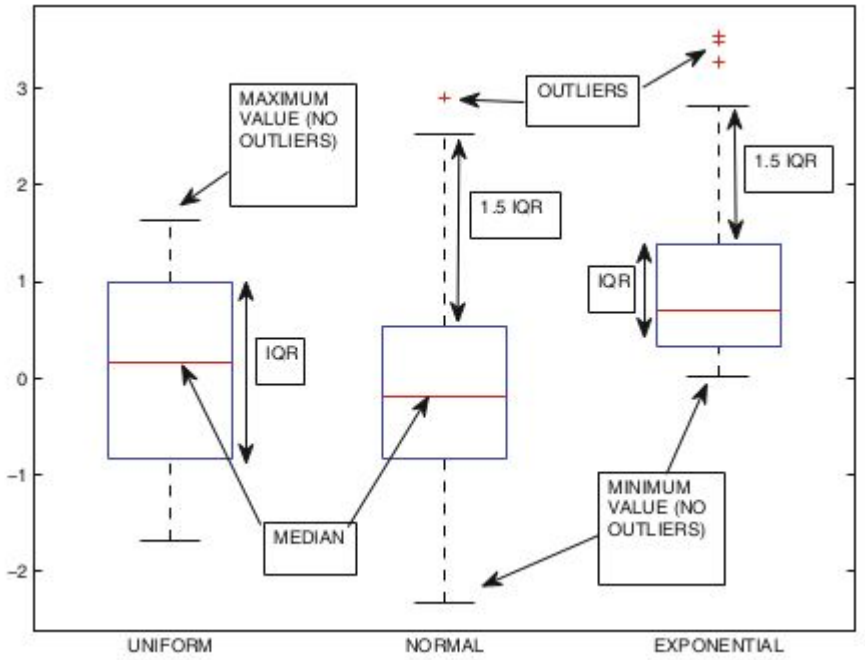
# Métodos Univariados

# Análisis de Box-Plot

Los Box-Plots permiten visualizar valores extremos univariados.

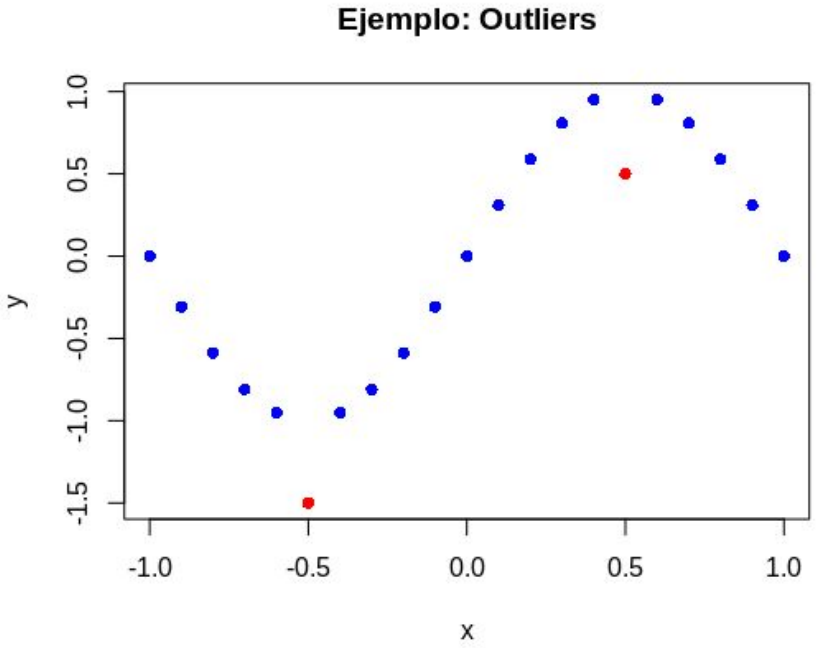
Las estadísticas de una distribución univariada se resumen en términos de cinco cantidades:

- Mínimo/máximo (bigotes)
- Primer y tercer cuantil (caja)
- Mediana (línea media de la caja)
- IQR = Q3 - Q1
- Generalmente la regla de decisión es **+/- 1.5*IQR**

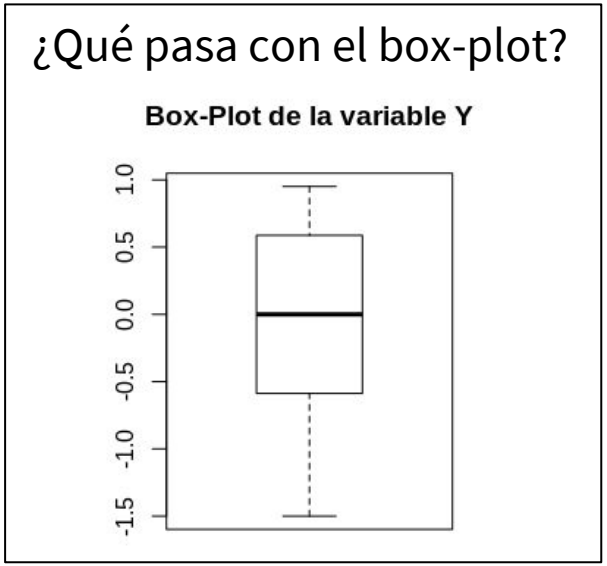


# Limitaciones del Box-Plot

En el scatter se observan dos valores atípicos.




¿Qué pasa con el box-plot?

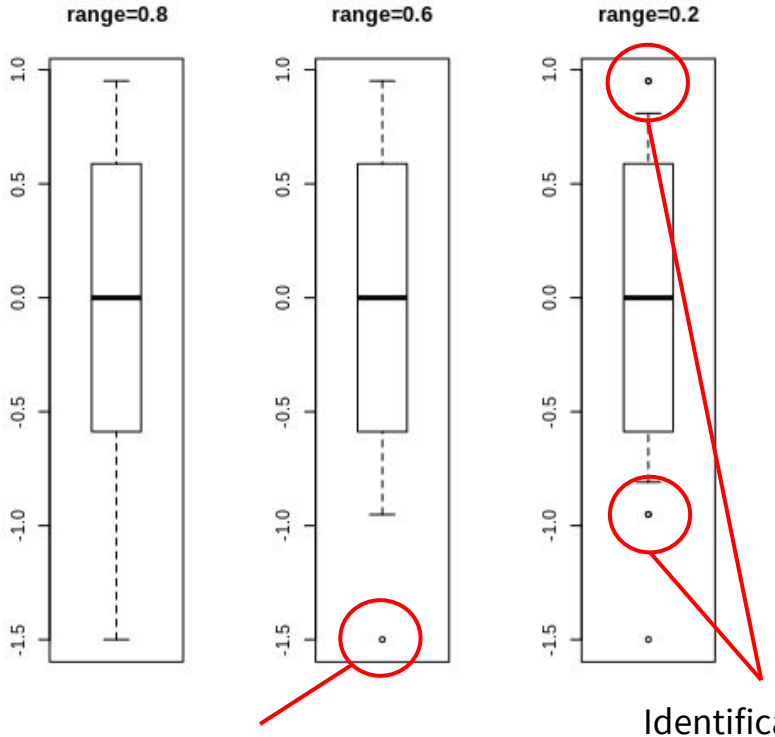


# Limitaciones del Box-Plot

Podemos intervenir modificando el largo de los bigotes del gráfico.

En R podemos hacerlo con el parámetro *range*.

También es posible modificar los cuantiles para regular el tamaño de las cajas.

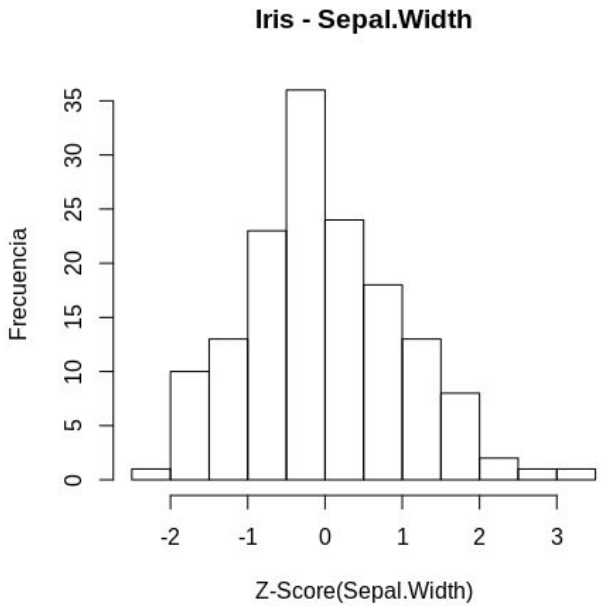


# Z-Score

Z-Score es una métrica que indica cuántas desviaciones estándar tiene una observación de la media muestral, asumiendo una distribución gaussiana.

$$z_i = \frac{x_i - \mu}{\sigma}$$

- Cuando calculamos Z-Score para cada muestra debemos fijar un umbral.
- Un buen umbral puede ser: 2.5, 3, 3.5 o más desviaciones estándar.



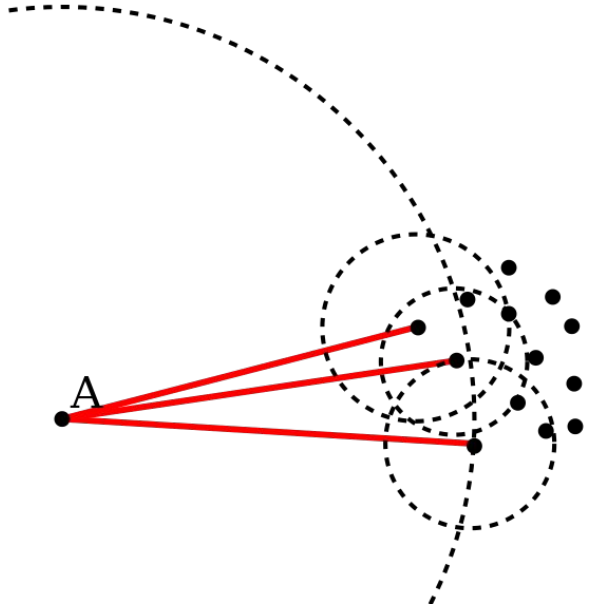
# Métodos Multivariados

# Local Outlier Factor (LOF)

El método del LOF valora puntos en un conjunto de datos multivariados cuyas filas se supone que se generan de forma independiente.

LOF es un método basado en densidad que utiliza la búsqueda de vecinos más cercanos.

El método calcula los *scores* para cada uno de los puntos a partir de la tasa promedio de densidad de los puntos vecinos con respecto a si mismo.



# Componentes del cálculo de LOF

K-Distancia de una observación ***o***, $dist_k(o)$: Es la distancia entre ***o*** y su ***k***-ésimo vecino más cercano (kNN)

Vecindad de ***o***, $N_k(o) = \{o' \mid o' \text{ en } D, dist(o, o') \leq dist_k(o)\}$

La distancia de vecindad son todas las distancias dentro del radio de $dist_k(o)$

Distancia esperada: de ***o'*** a ***o***: $reachdist(o \leftarrow o') = max(dist_k(o), dist(o, o'))$

- Densidad Local Esperada de ***o***:
  *(Local Reachability Density)*

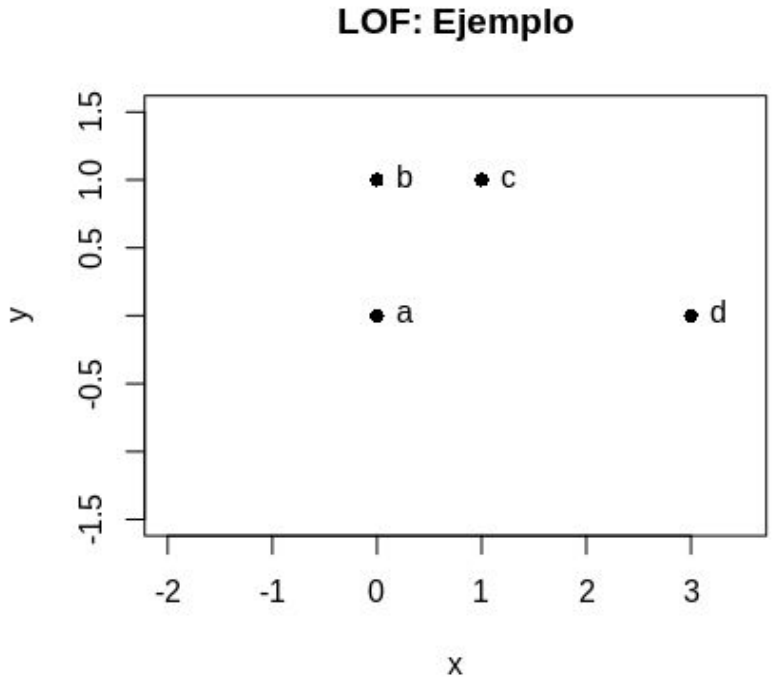
$$lrd_k(o) = \frac{\|N_k(o)\|}{\sum_{o' \in N_k(o)} reachdist_k(o' \leftarrow o)}$$

$$LOF_k(o) = \frac{\sum_{o' \in N_k(o)} \frac{lrd_k(o')}{lrd_k(o)}}{\|N_k(o)\|} = \sum_{o' \in N_k(o)} lrd_k(o') \cdot \sum_{o' \in N_k(o)} reachdist_k(o' \leftarrow o)$$

# LOF: Ejemplo

Dado el siguiente conjunto de datos:

| | x | y |
|---|---|---|
| **a** | 0 | 0 |
| **b** | 0 | 1 |
| **c** | 1 | 1 |
| **d** | 3 | 0 |




**Paso 1**: Calcular la matriz de distancias.

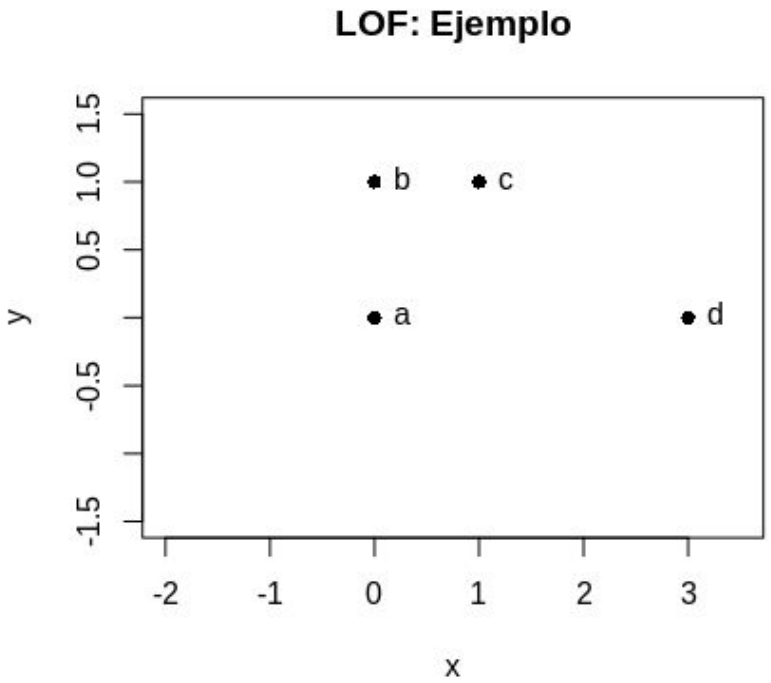
| | a | b | c |
|---|---|---|---|
| **b** | 1 | | |
| **c** | 2 | 1 | |
| **d** | 3 | 4 | 3 |

Para el ejemplo utilizamos distancia de Manhattan:

$$d_1(\mathbf{p}, \mathbf{q}) = \|\mathbf{p} - \mathbf{q}\|_1 = \sum_{i=1}^{n} |p_i - q_i|$$

# LOF: Ejemplo

**Paso 2**: Calculamos las distancias al k-ésimo vecino más cercano. Para este ejemplo **k=2.**




$Dist_2(a) = dist(a, c) = 2.$

La distancia del k=2 vecino más cercano es la distancia de a → c

Recordemos la matriz de distancias para a:

El 2do NN es **c**

| | **a** | **b** | **c** |
|---|---|---|---|
| **b** | 1 | | |
| **c** | 2 | 1 | |
| **d** | 3 | 4 | 3 |

$Dist_2(b) = dist(b, a) = 1$ También puede estar c

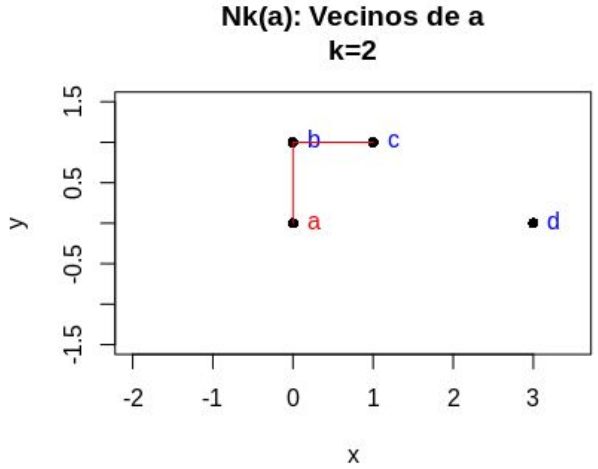
$Dist_2(c) = dist(c, a) = 2$

$Dist_2(d) = dist(d, a) = 3$ También puede estar c

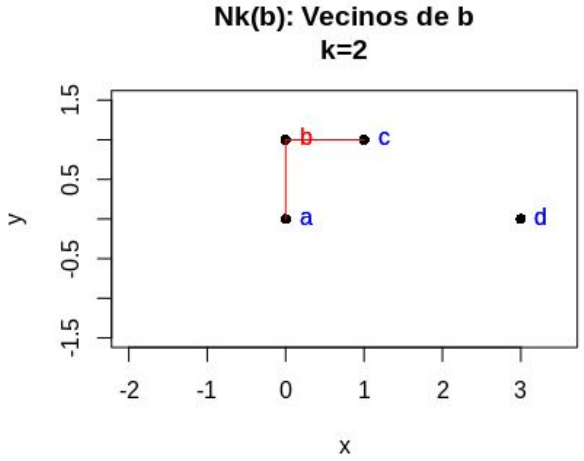
# LOF: Ejemplo

**Paso 3**: Calculamos todas las distancias de la vecindad: $\mathbf{N_k(o)}$

$$\mathbf{N_k(o) = \{o' | \, o' \text{ in } D, \text{dist}(o, o') \leq \text{dist k } (o)\}}$$




$N_2(a) = \{b, c\}$

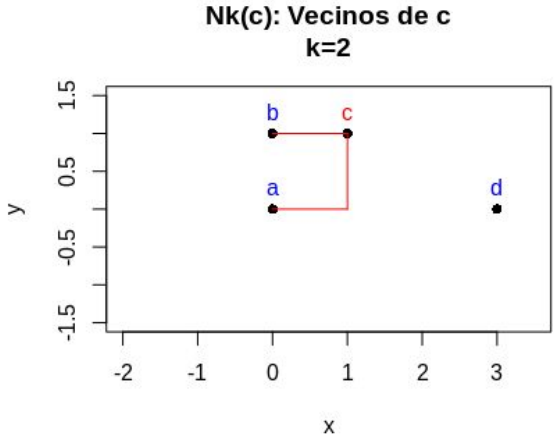



$N_2(b) = \{a, c\}$

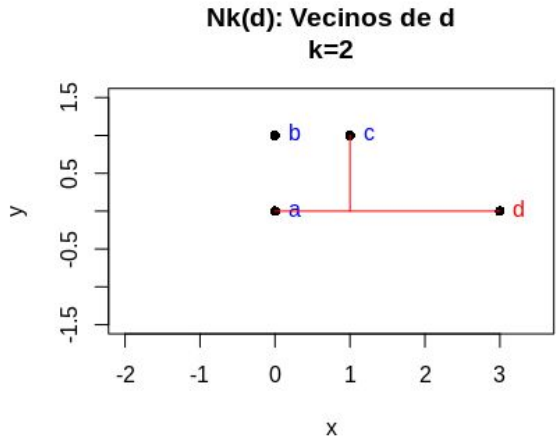
# LOF: Ejemplo

**Paso 3**: Calculamos todas las distancias de la vecindad: $\mathbf{N_k(o)}$

$$\mathbf{N_k(o) = \{o' | \ o' \text{ in } D, \text{dist}(o, o') \leq \text{dist k } (o)\}}$$




$N_2(c) = \{b, a\}$




$N_2(d) = \{a, c\}$

# LOF: Ejemplo

**Paso 4**: Calculamos todos los **$LRD_k(o)$**

$$lrd_k(o) = \frac{\|N_k(o)\|}{\sum_{o' \in N_k(o)} reachdist_k(o' \leftarrow o)}$$

$$reachdist_k(o \leftarrow o') = \max\{dist_k(o), dist(o, o')\}$$

**|| $N_k(o)$ ||** es la cantidad de objetos en **$N_k(o)$**,

por ejemplo: $\|N_2(a)\| = \|\{b, c\}\| = 2$

$$LDR_k(a) = \frac{\|N_2(a)\|}{reachdist_2(b \leftarrow a) + reachdist_2(c \leftarrow a)}$$

$$reachdist_2(b \leftarrow a) = max[dist_2(b), dist(b, a)] = max[1, 1] = 1$$
$$reachdist_2(c \leftarrow a) = max[dist_2(c), dist(c, a)] = max[2, 2] = 2$$

$$LDR_2(a) = \frac{\|N_2(a)\|}{reachdist_2(b \leftarrow a) + reachdist_2(c \leftarrow a)} = \frac{2}{1+2} = 0.667$$

$$LDR_2(b) = \frac{\|N_2(b)\|}{reachdist_2(a \leftarrow b) + reachdist_2(c \leftarrow b)} = \frac{2}{2+2} = 0.5$$

$$LDR_2(c) = \frac{\|N_2(c)\|}{reachdist_2(a \leftarrow c) + reachdist_2(b \leftarrow c)} = \frac{2}{1+2} = 0.667$$

$$LDR_2(d) = \frac{\|N_2(d)\|}{reachdist_2(a \leftarrow d) + reachdist_2(c \leftarrow d)} = \frac{2}{3+3} = 0.33$$

# LOF: Ejemplo

**Paso 5**: Calcular **$LOF_k(o)$**

$$LOF_k(o) = \frac{\sum_{o' \in N_k(o)} \frac{lrd_k(o')}{lrd_k(o)}}{\|N_k(o)\|} = \sum_{o' \in N_k(o)} lrd_k(o') \cdot \sum_{o' \in N_k(o)} reachdist_k(o' \leftarrow o)$$

$$LOF_2(a) = (lrd_2(b) + lrd_2(c)) * (reachdist_2(b \leftarrow a) + reachdist_2(c \leftarrow a)) = (0.5 + 0.667) * (1 + 2) = 3.501$$

$$LOF_2(b) = (lrd_2(a) + lrd_2(c)) * (reachdist_2(a \leftarrow b) + reachdist_2(c \leftarrow b)) = (0.667 + 0.667) * (2 + 2) = 5.336$$

$$LOF_2(c) = (lrd_2(b) + lrd_2(a)) * (reachdist_2(b \leftarrow c) + reachdist_2(a \leftarrow c)) = (0.5 + 0.667) * (1 + 2) = 3.501$$

$$LOF_2(d) = (lrd_2(a) + lrd_2(c)) * (reachdist_2(a \leftarrow d) + reachdist_2(c \leftarrow d)) = (0.667 + 0.667) * (3 + 3) = 8.004$$

# LOF: Ejemplo

**Paso 6**: Ordenamos todos los valores de $\mathbf{LOF_k(o)}$

$LOF_2(d) = \mathbf{8.004}$

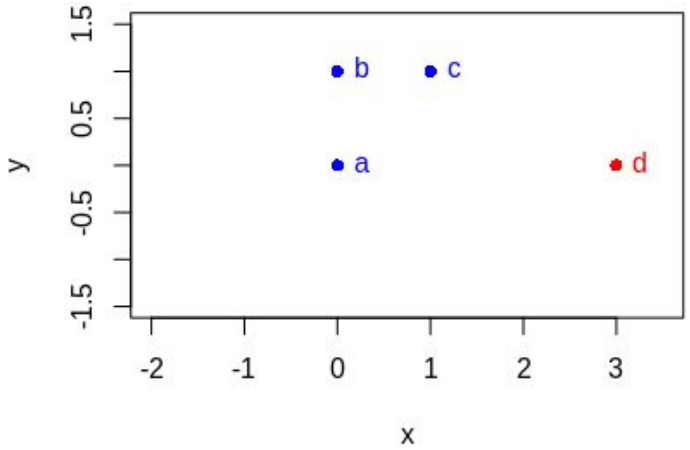
$LOF_2(b) = 5.336$

$LOF_2(a) = 3.501$

$LOF_2(c) = 3.501$

Una vez que ordenamos tenemos que definir un umbral de corte.

En este caso el punto **d** es un claro valor atípico.



# Bibliografía


- Jiawei Han, Micheline Kamber, Jian Pei. 2012. Tercera edición. Data Mining: Concepts and Techniques. Cap. 2 y Cap. 3
- http://www.cse.ust.hk/~leichen/courses/comp5331/lectures/LOF_Example.pdf