Exploración de datos

Dr. Raimundo Sánchez

raimundo.sanchez@uai.cl

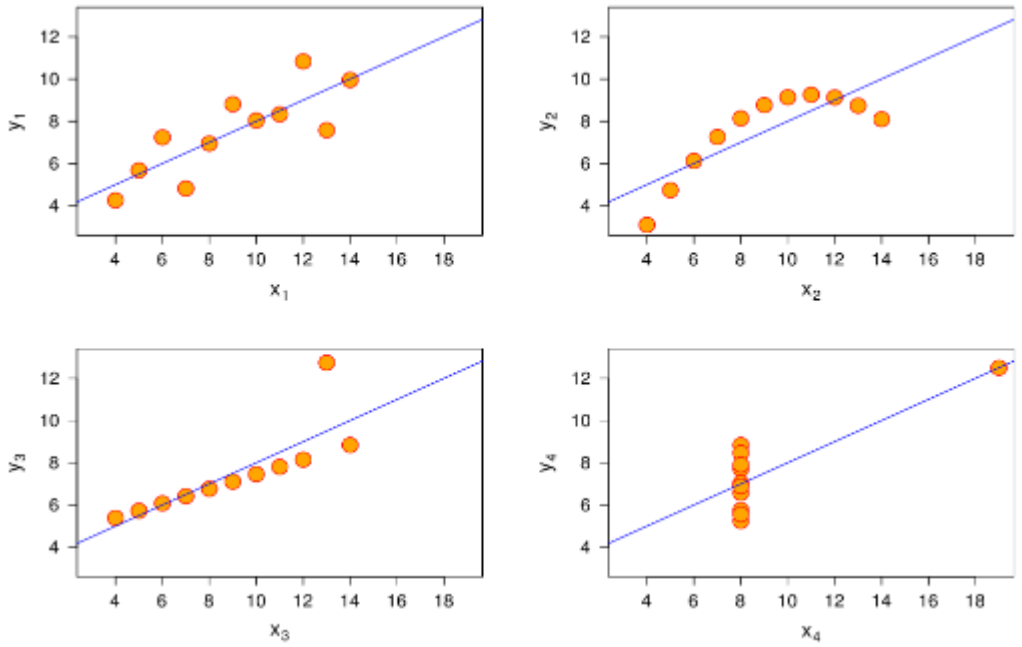
@raimun2

Análisis exploratorio de datos (EDA)

Análisis del conjunto de datos para resumir sus principales características, mediante métodos estadísticos y visuales.

Objetivos:

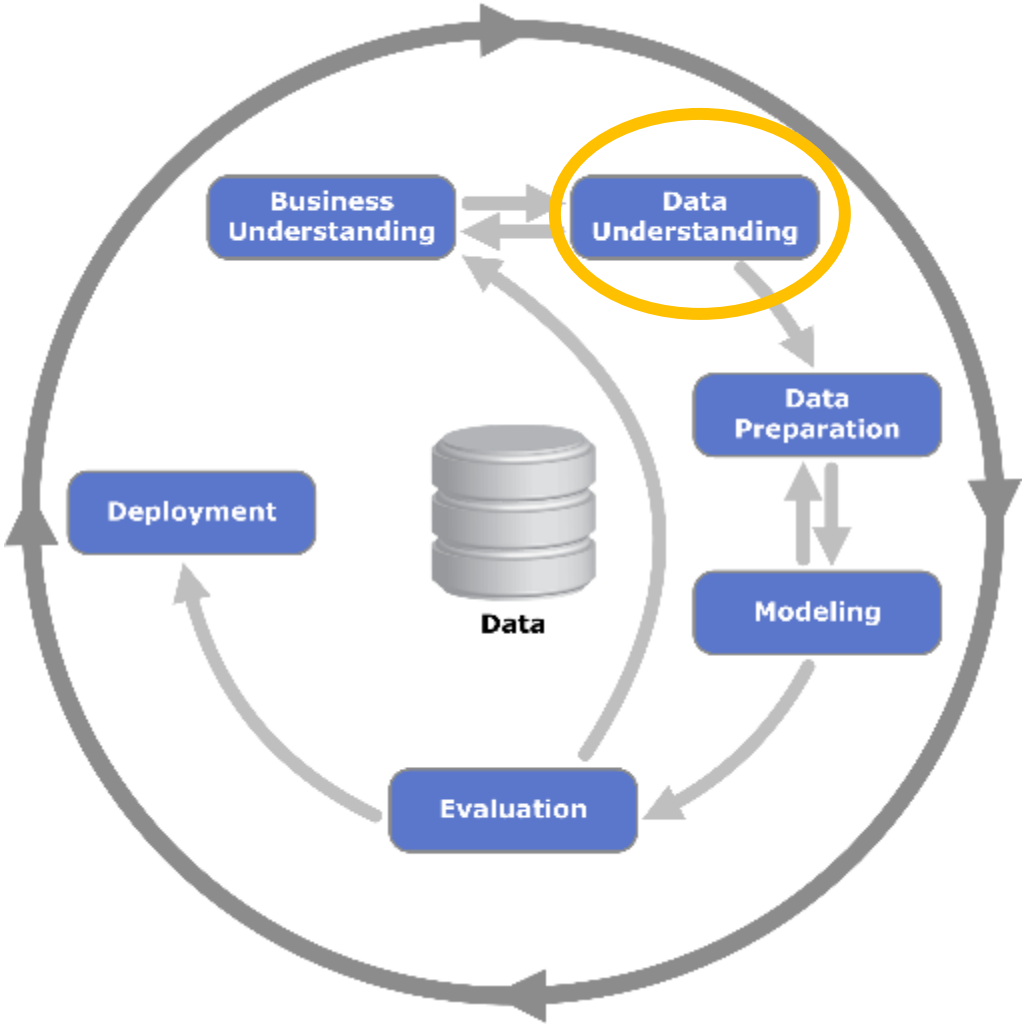
- Descubrir la estructura subyacente de los datos
- Identificar variables relevantes
- Detectar valores atípicos y anomalías
- Validar supuestos
- Generar hipótesis a partir de los datos



Cuarteto de Anscombe

Propiedad	Valor
Media de cada una de las variables *x*	9.0
Varianza de cada una de las variables *x*	11.0
Media de cada una de las variables *y*	7.5
Varianza de cada una de las variables *y*	4.12
Correlación entre cada una de las variables *x* e *y*	0.816
Recta de regresión	y=3+0.5x

CRISP DM


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

Técnicas de EDA

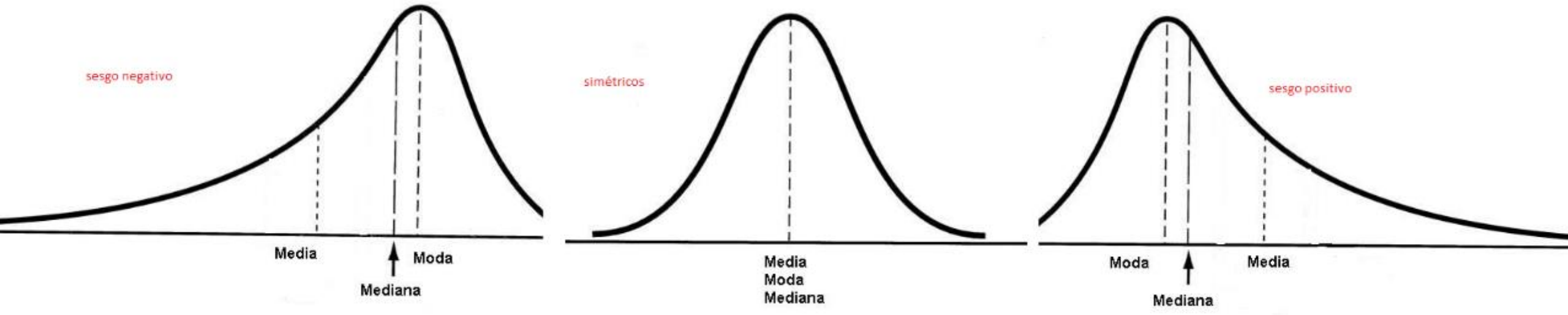
- **Estadística descriptiva:** Métricas de tendencia central, dispersión y variabilidad.
- **Test estadísticos:** Permiten validar estadísticamente el comportamiento de las variables.
- **Análisis multivariable**: Permite identificar variables redundantes, observaciones similares y datos atípicos.
- **Visualización**: Uso de métodos gráficos para inspeccionar visualmente los datos.

Estadística descriptiva

Medidas de distribución

- Media
- Moda
- Mediana
- Cuantiles
- Varianza
- Desviación estándar

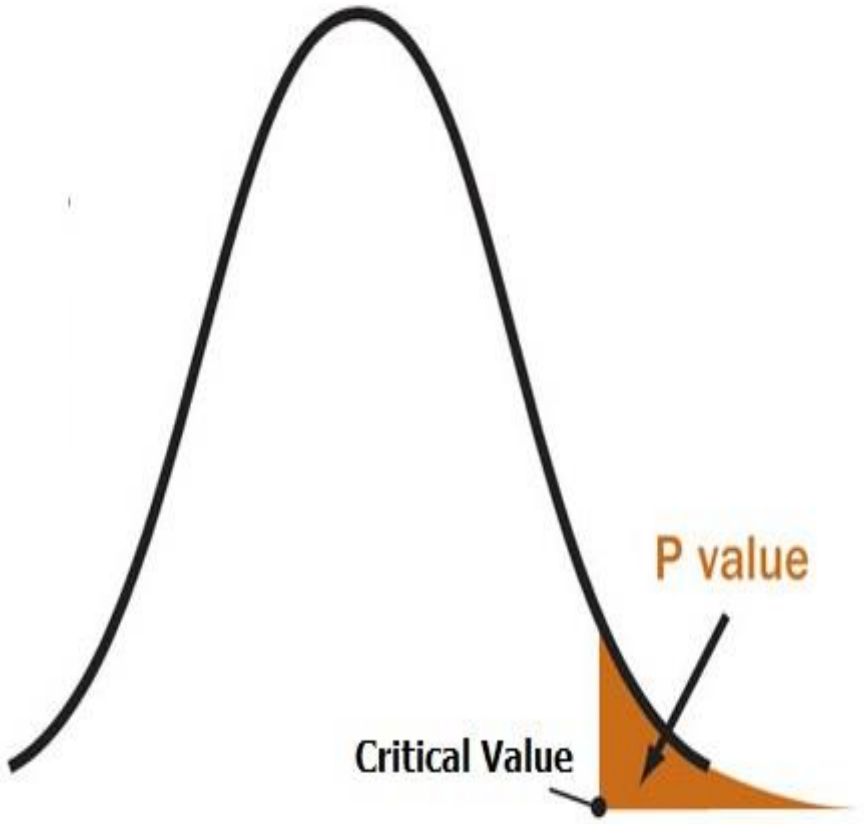
- Coeficiente de variación
- Rango
- Rango inter cuartil
- Sesgo
- Entropía



Test estadísticos

Se asumen conocidos para este curso

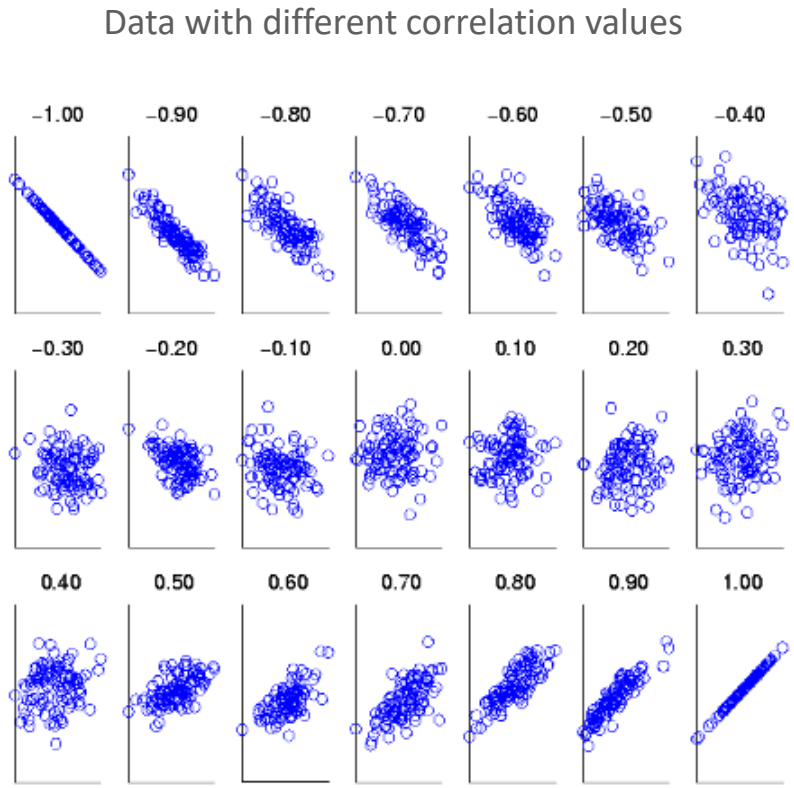
- Intervalos de confianza
- Test de hipótesis sobre la media y varianza
- Test de distribuciones (Kolmogorov-Smirnov)
- One-way ANOVA
- Kruskal-Wallis.



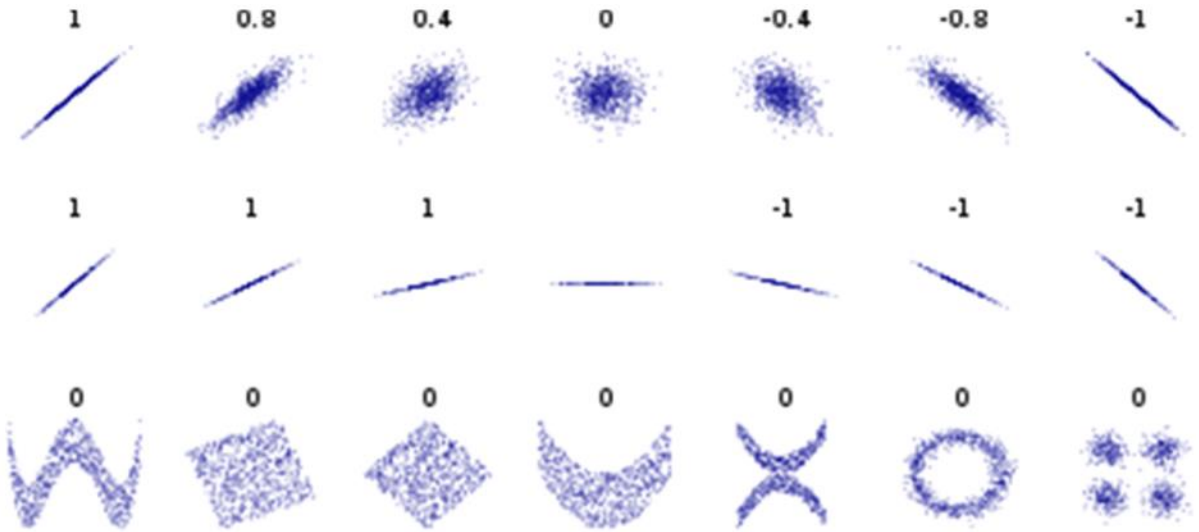
Análisis multivariable

Correlación de Pearson

- Correlación de Pearson mide la relación lineal entre atributos
- **CUIDADO**: No es lo mismo que una regresión lineal



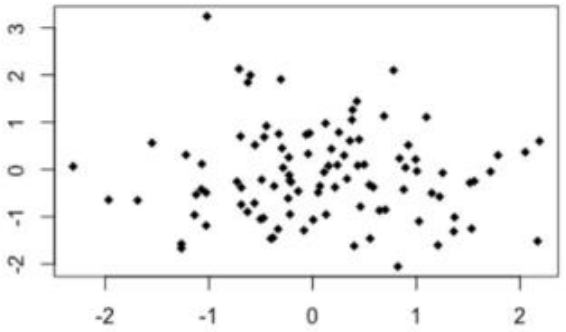
Correlación



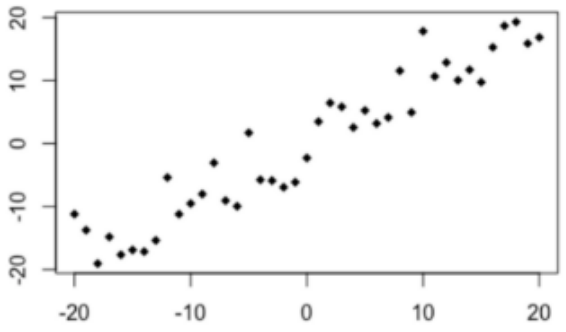
No todas las relaciones entre variables son lineales

Correlación de distancia

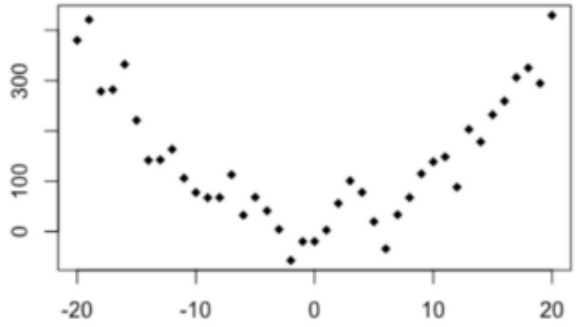
Correlación de distancia mide la relación no lineal (y lineal) entre atributos.



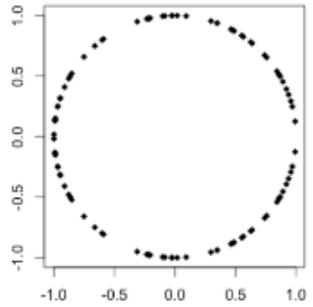
- Pearson's r = - 0.05
- Distance Correlation = 0.157
- MIC = 0.097



- Pearson's r =+0.95
- Distance Correlation = 0.95
- MIC = 0.89



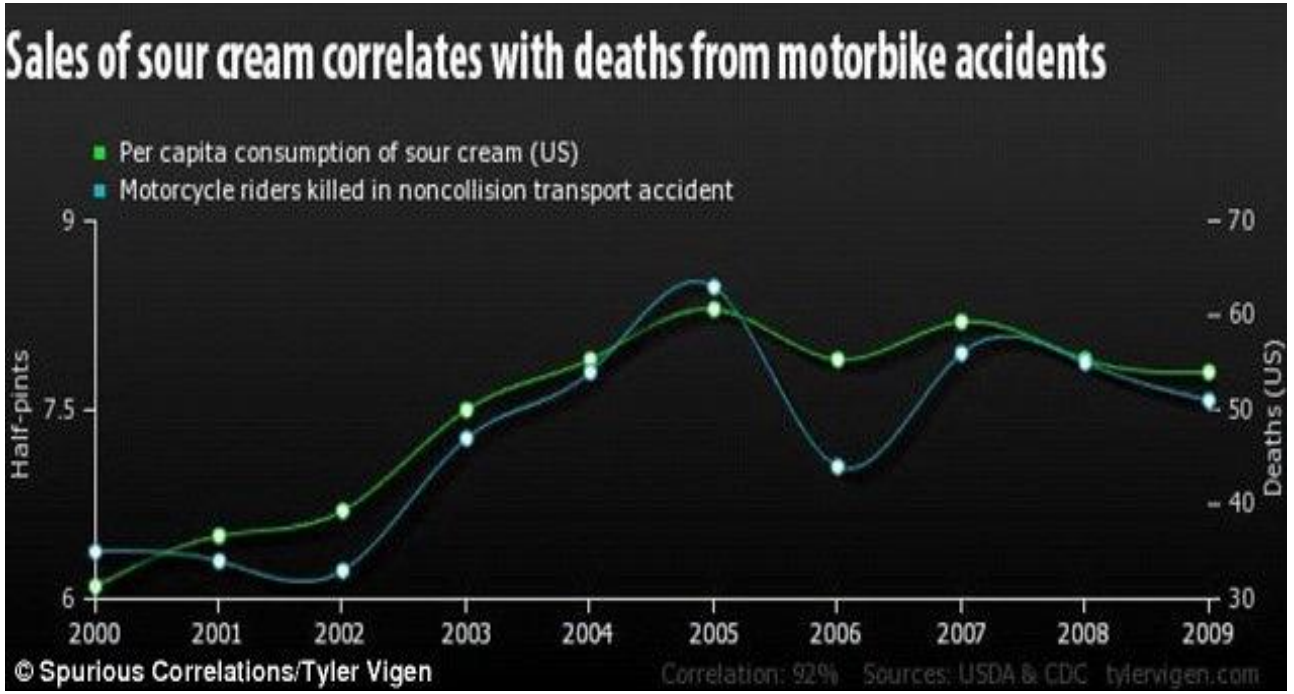
- Pearson's r =+0.003
- Distance Correlation = 0.474
- MIC = 0.594



- Pearson's r < 0.001
- Distance Correlation = 0.234
- MIC = 0.218

Otras correlaciones

- Maximal Information Coefficient “MIC”
- Correlación de Spearman
- Tau de Kendall



Similitud

Similitud

- Mide lo parecidos que son dos entidades.
- Es más alto cuando entidades son más parecidas.

Disimilitud

- Mide cuan diferentes son dos entidades
- Más bajo cuando los objetos son más parecidos

Tipo de Atributo	Disimilitud	Similitud
Nominal	$d = \begin{cases} 0 & \text{if } p = q \\ 1 & \text{if } p \neq q \end{cases}$	$s = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q \end{cases}$
Ordinal	$d = \frac{\lvert p-q \rvert}{n-1}$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - \frac{\lvert p-q \rvert}{n-1}$
Intervalo o Ratio	$d = \lvert p - q \rvert$	$s = -d$, $s = \frac{1}{1+d}$ or $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

p y q son los valores de un atributo para dos entidades

Similitud de datos categóricos

- Sean p y q vectores m-dimensionales con solo atributos categóricos.
- Sea pk(x) la proporción de registros en la que el k-ésimo atributo adquiere el valor x en el conjunto de datos.

- Para calcular las similitudes entre estos dos vectores, sumamos la similitud S de cada atributo i

$$Sim(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{m} S(p_i, q_i)$$

La medida de **superposición** es *S*(*pi*,*qi*)=1 si pi = qi y 0 de otra manera.

La **frecuencia de ocurrencia inversa** pondera la similitud entre los atributos coincidentes.

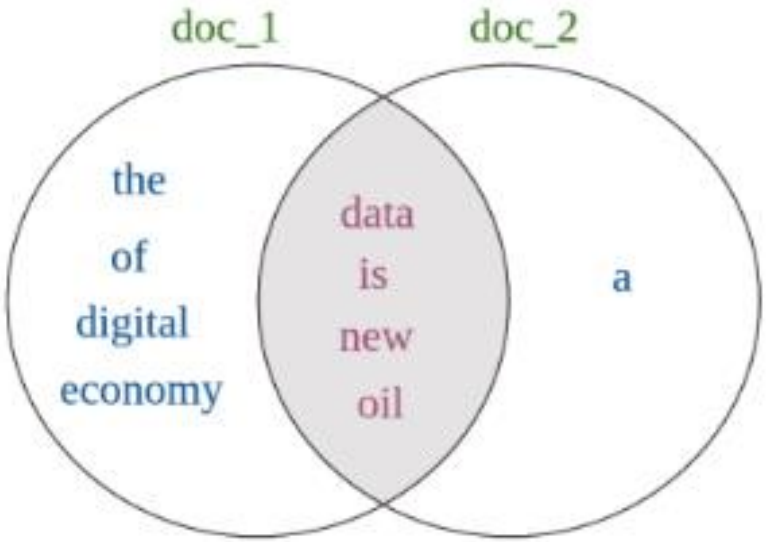
$$S(\mathrm{p}_i, \mathrm{q}_i) = 1/p_k(\mathrm{p}_i)^2 \quad \text{if } \mathrm{p}_i = \mathrm{q}_i$$ **sino $S(p_i,q_i) = 0$**

La medida de **Goodall** da mayor peso a valores poco frecuentes.

$$S(\mathrm{p}_i, \mathrm{q}_i) = 1 - p_k(\mathrm{p}_i)^2 \quad \text{if } \mathrm{p}_i = \mathrm{q}_i$$ **sino $S(p_i,q_i) = 0$**

Similitud de vectores binarios

- Sean p y q vectores solo con atributos binarios (Word embedding)
- Para calcular las similitudes entre estos dos vectores, utilizamos las siguientes métricas:
  - M01 = el número de atributos donde p era 0 y q era 1
  - M10 = el número de atributos donde p era 1 y q era 0
  - M00 = el número de atributos donde p era 0 y q era 0
  - M11 = el número de atributos donde p era 1 y q era 1



Coeficiente simple de coincidencia (SMC):
número de coincidencias / número de atributos
(M00 + M11) / (M00 + M01 + M10 + M111)

Coeficiente de Jaccard (J):
número de coincidencias 11 / número de atributos no 00
(M11) / (M01 + M10 + M11)

Similitud de vectores binarios: ejemplo

p = “El cerro esta muy nevado hoy”
q = “Me gustaría ir al cerro hoy”
w = “Al gato le gustaría comer”

→

D1	el
D2	cerro
D3	esta
D4	muy
D5	nevado
D6	hoy
D7	me
D8	gustaría
D9	ir
D10	al
D11	gato
D12	le
D13	comer

→

Dimensiones:	1	2	3	4	5	6	7	8	9	10	11	12	13
p :	1	1	1	1	1	1	0	0	0	0	0	0	0
q :	0	1	0	0	0	1	1	1	1	1	0	0	0
w :	0	0	0	0	0	0	0	1	0	1	1	1	1

Binarización

M01 = 4 (el número de atributos donde p era 0 y q era 1)
M10 = 4 (el número de atributos donde p era 1 y q era 0)
M00 = 3 (el número de atributos donde p era 0 y q era 0)
M11 = 2 (el número de atributos donde p era 1 y q era 1)

$SMC(p,q) = (M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00}) = (2+0) / (4+4+2+3) = 0.153$

$J(p,q) = (M_{11}) / (M_{01} + M_{10} + M_{11}) = 2 / (4 + 4 + 2) = 0.2$

Similitud coseno

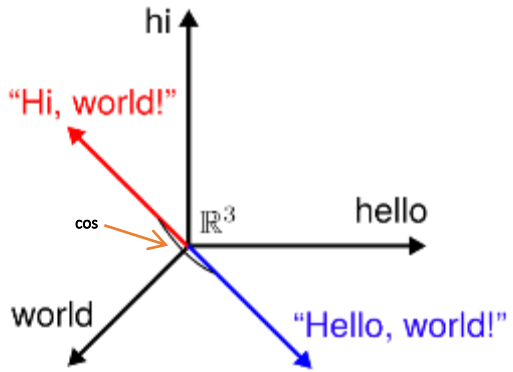
- Calcula el ángulo entre los dos documentos, lo que es insensible a la longitud del documento.

- Si a y b son dos vectores de documento se calcula como:

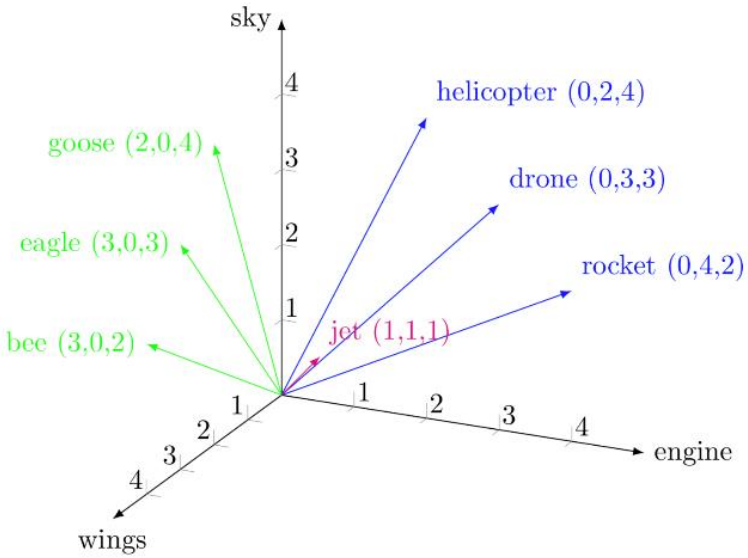
$$\cos\theta = \frac{\vec{a}\cdot\vec{b}}{\|\vec{a}\|\|\vec{b}\|}$$

$$\|\vec{a}\| = \sqrt{a_1^2 + a_2^2 + a_3^2 + \cdots + a_n^2}$$

$$\|\vec{b}\| = \sqrt{b_1^2 + b_2^2 + b_3^2 + \cdots + b_n^2}$$



Cosine Similarity



Distancia

- Una métrica que mide la distancia entre un par de entidades
- Dados los dos puntos x e y, una función métrica o de distancia debe cumplir las siguientes condiciones
  - no negatividad => d(x, y) > 0
  - identidad => d(x, y) = 0 <=> x=y
  - simetría => d(x, y) = d(y, x)
  - desigualdad triangular => d(x, z) <= d(x, y) + d(y, z)

Distancia de Minkowski

- Generalización de la distancia euclidiana.
- Sean que p y q vectores m-dimensionales

$$d(p,q) = \left(\sum_{k=1}^{m} |p_k - q_k|^r \right)^{1/r}$$

Blue: Eclidean
Red: Manhattan
Orange: Chebyshev

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

Para r = 2 : Distancia euclidiana: $d(p,q) = \sqrt{\sum_{k=1}^{m} (p_k - q_k)^2}$

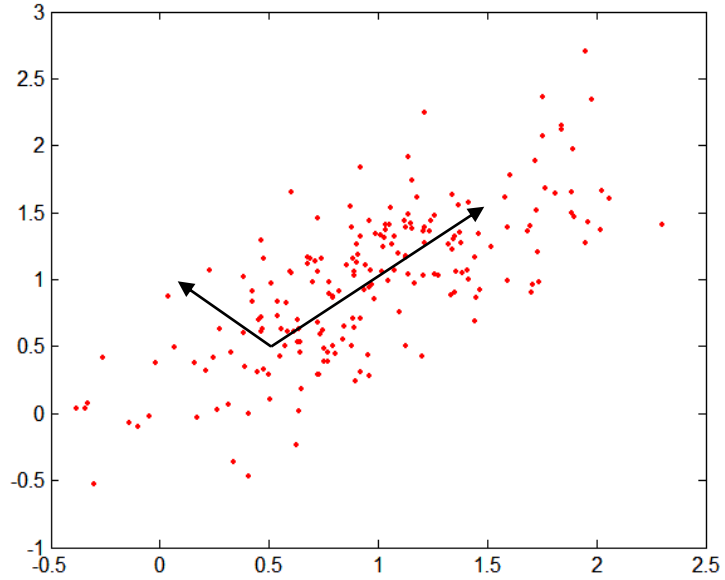
Para r =∞ : Distancia Chebyshev (distancia suprema) $\lim_{k\to\infty} \left(\sum_{i=1}^{n} |p_i - q_i|^k \right)^{1/k}$ $D_{\text{Chebyshov}}(p,q) := \max_i(|p_i - q_i|).$

Distancia de Mahalanobis

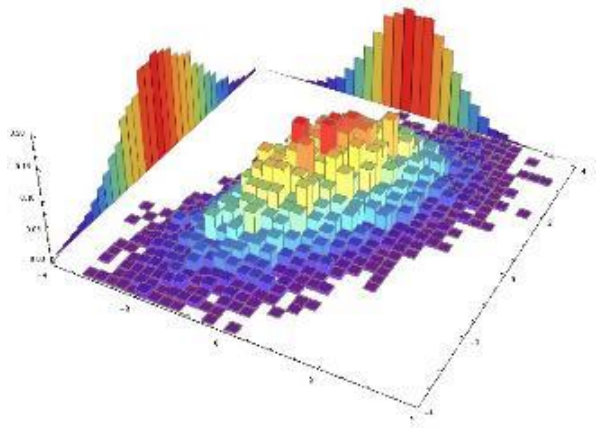
- Considera la varianza de los datos para calcular la distancia.

$$d(p,q) = \sqrt{(p-q)^T \Sigma^{-1} (p-q)}$$

donde ∑ es la matriz de covarianza de los datos de entrada.



- Usada para calcular distancias cuando hay atributos correlacionados.



Distancia del coseno

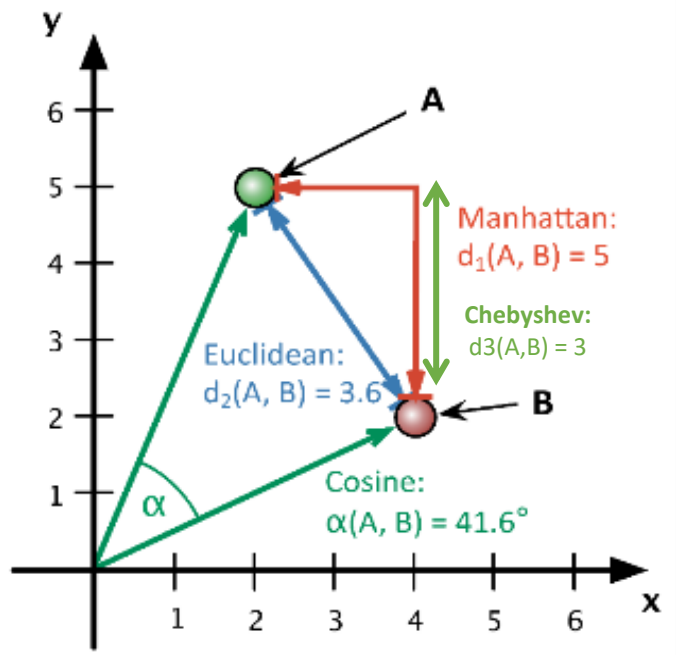
- La distancia del coseno es el espacio de complemento positivo de similitud coseno.
- La similitud coseno es una medida de similitud entre dos vectores no nulos de un espacio interior del producto.

- Cuando los elementos vectoriales pueden ser positivos o negativos:

$$\text{angular distance} = \frac{\cos^{-1}(\text{cosine similarity})}{\pi}$$

- Cuando los elementos vectoriales siempre son positivos:

$$\text{angular distance} = \frac{2 \cdot \cos^{-1}(\text{cosine similarity})}{\pi}$$

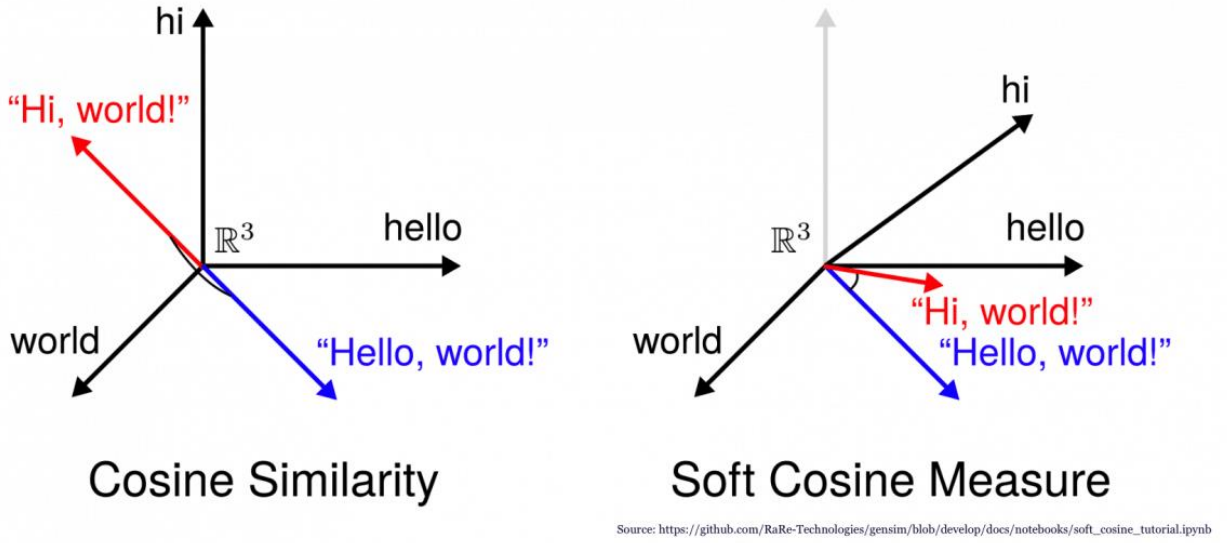


Distancia suave del coseno

- Distancia suave del coseno considera la similaridad de los atributos
- Comúnmente usado en procesamiento de lenguaje natural (NLP) para representar que conceptos cercanos (como “manteca” y “mantequilla”)

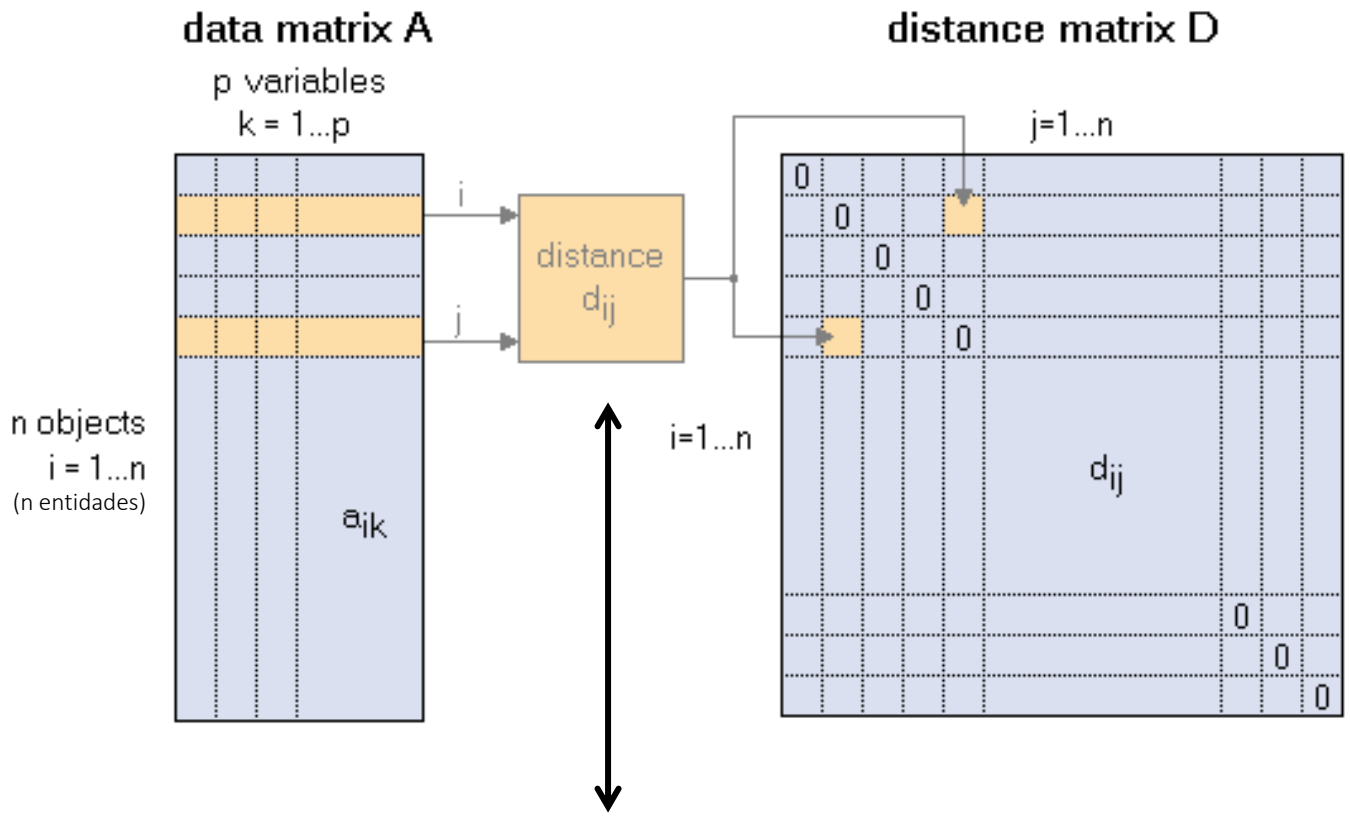
$$\text{soft\_cosine}_1(a, b) = \frac{\sum_{i,j}^{N} s_{ij} a_i b_j}{\sqrt{\sum_{i,j}^{N} s_{ij} a_i a_j} \sqrt{\sum_{i,j}^{N} s_{ij} b_i b_j}},$$

where $s_{ij} = \text{similarity}(\text{feature}_i, \text{feature}_j)$.



Source: https://github.com/RaRe-Technologies/gensim/blob/develop/docs/notebooks/soft_cosine_tutorial.ipynb

Matriz de Distancias



(Euclidean, Manhattan, Chebyshev, Mahalanobis, Cosine, Haversine, etc.)

Data Viz

Visualización de datos

- La visualización de datos es la presentación de datos en formato pictórico o gráfico
- Permite inspeccionar visualmente los datos para apoyar la comprensión de datos complejos
- Visualización debe ser auto explicativa
- Puede ser difícil de aplicar si el tamaño de los datos es grande o el numero de dimensiones es muy alto.

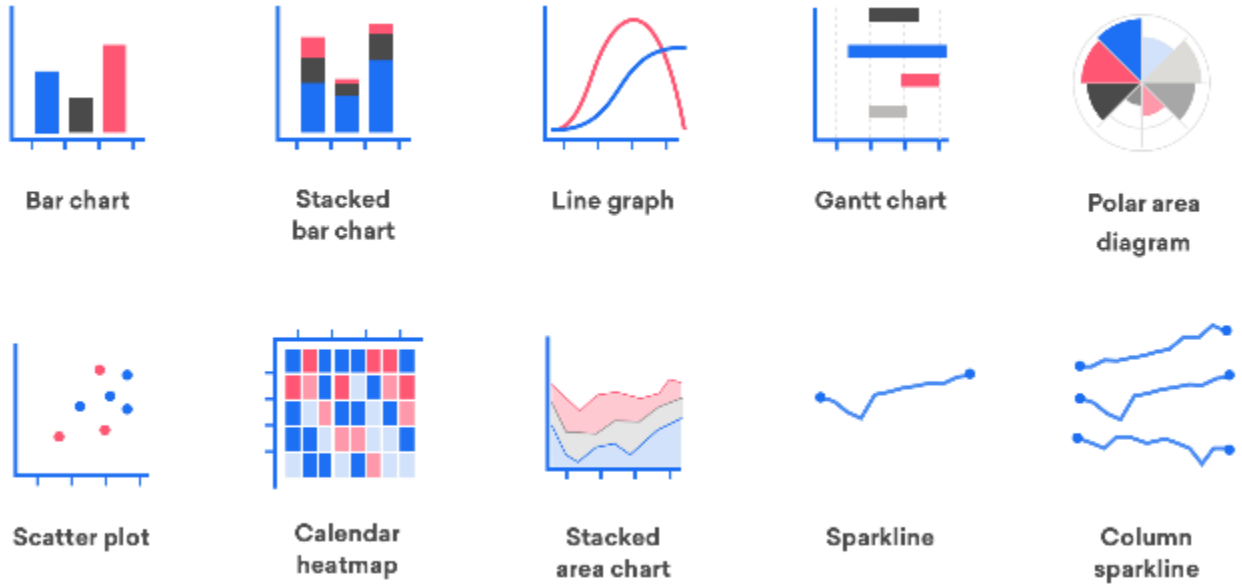


Chart Suggestions—A Thought-Starter

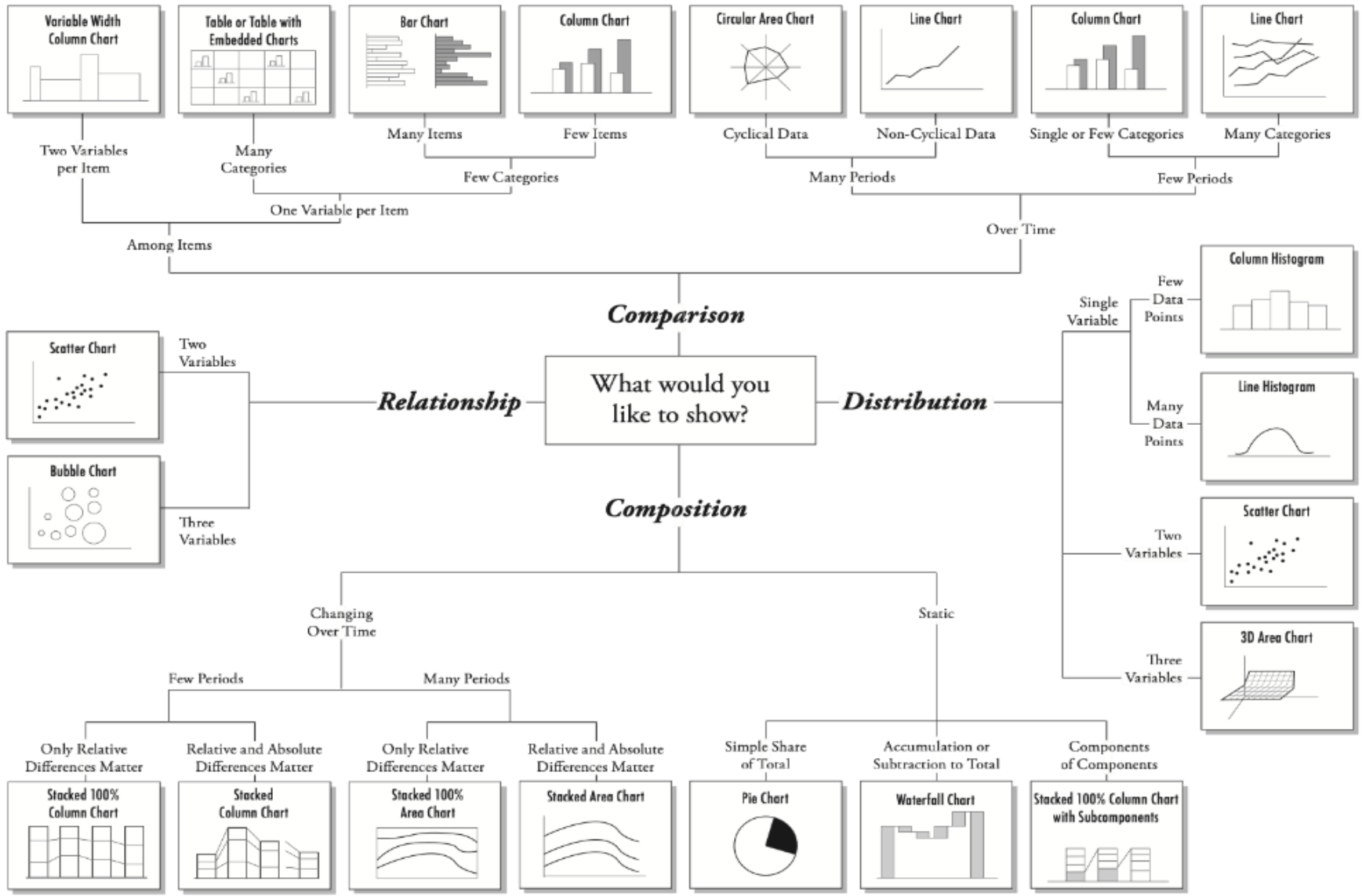
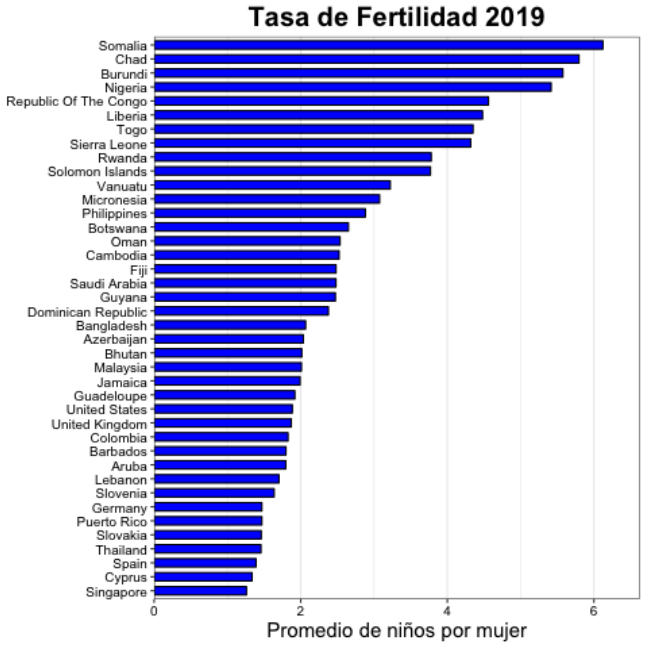
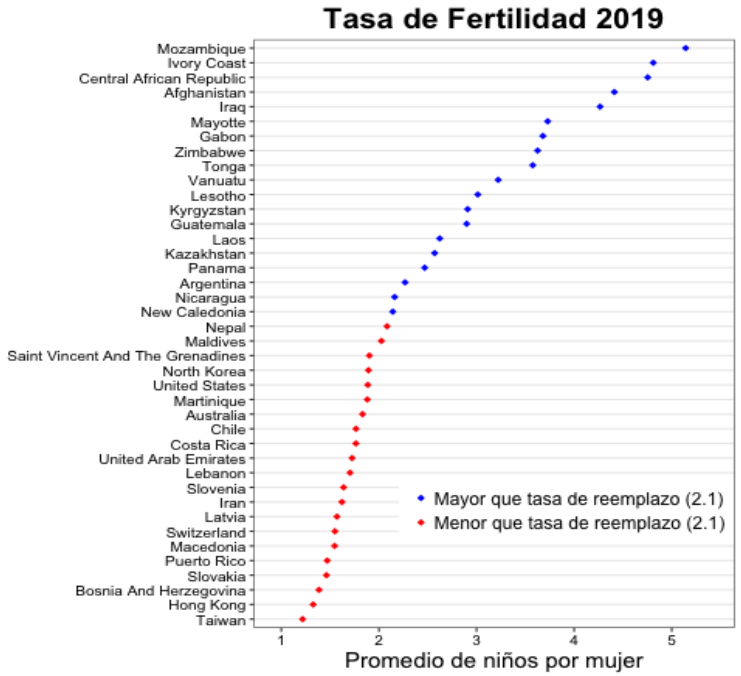


Gráfico de Barras / Puntos de Cleveland

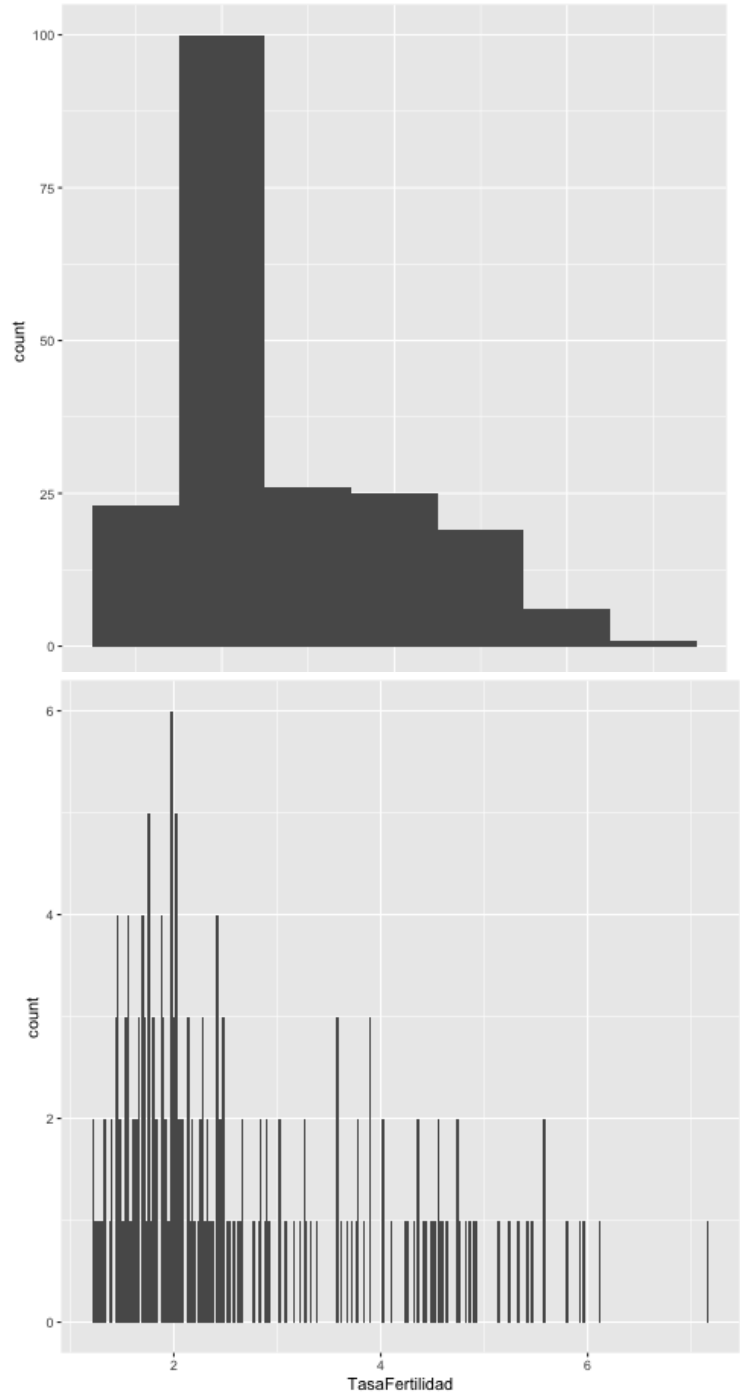
- Cada punto de datos se traza con una altura igual a su valor.
- El valor cero debe estar presente, de lo contrario la escala puede verse afectada.
- Ordene siempre los valores de menor a más alto o viceversa.





Histogramas

- Gráfico más común para datos univariados .
- Divide el rango de datos en contenedores y mide la frecuencia en cada rango.
- Muestra gráficamente el centro, moda, rango, sesgo y datos atípicos de una distribución.
- El número de contenedores cambia la percepción de la distribución.
- Número de contendores se define heurísticamente.



Gráficos de densidad de núcleo

- Método del núcleo utiliza una función no negativa que se integra a uno y tiene media cero para aproximar la distribución.
- Estima parámetros de densidad en base a un ancho de banda determinado.

La densidad estimada es:

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{x - x(i)}{h}\right)$$

Donde

- K : Función de núcleo o Kernel (uniforme, triangular, Epanechnikov, Gaussiana, etc.)
- h : Ancho de banda (análogo a tamaño de contenedor)

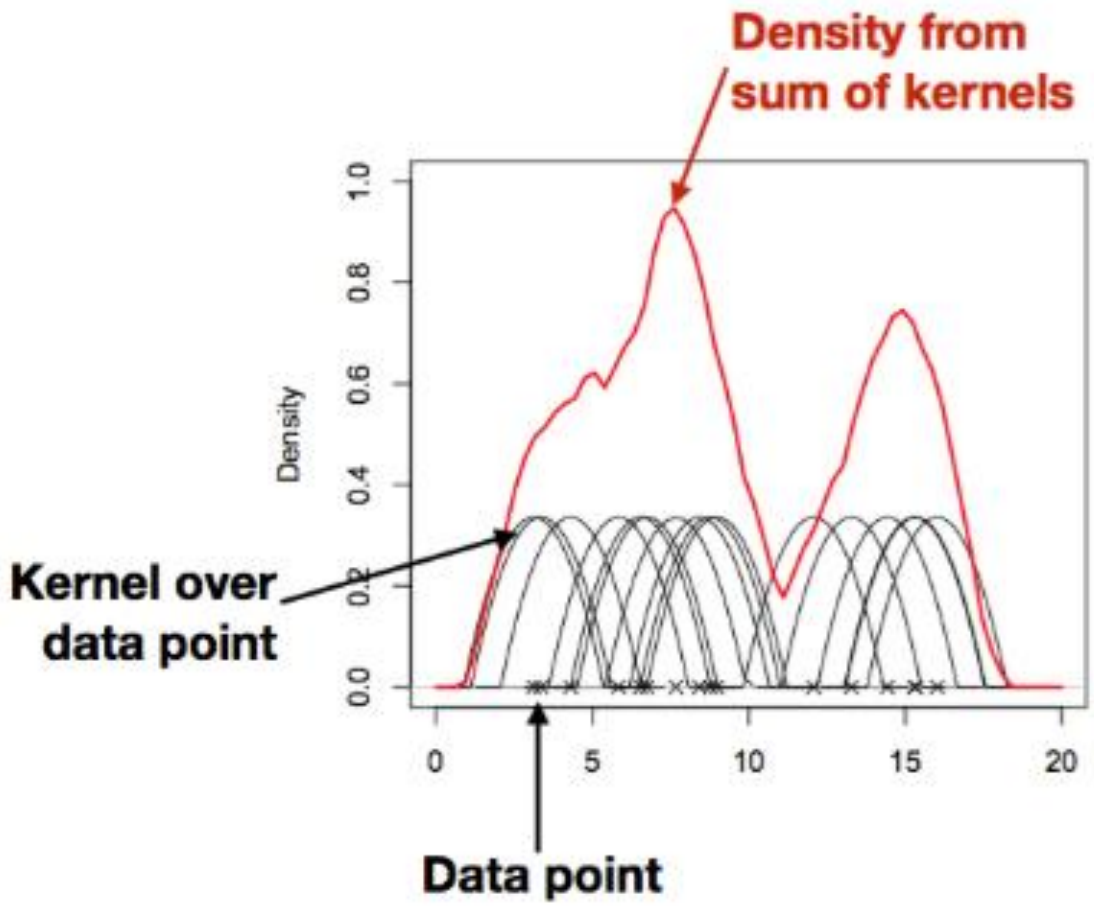


Diagrama de caja (y bigote)

- Muestra la relación entre variables discretas y continuas
- Calcula los cuartiles y rango de valores una variable
- Los bigotes máximos y mínimos se extienden a los puntos de datos más extremos que el algoritmo considera que no son valores atípicos

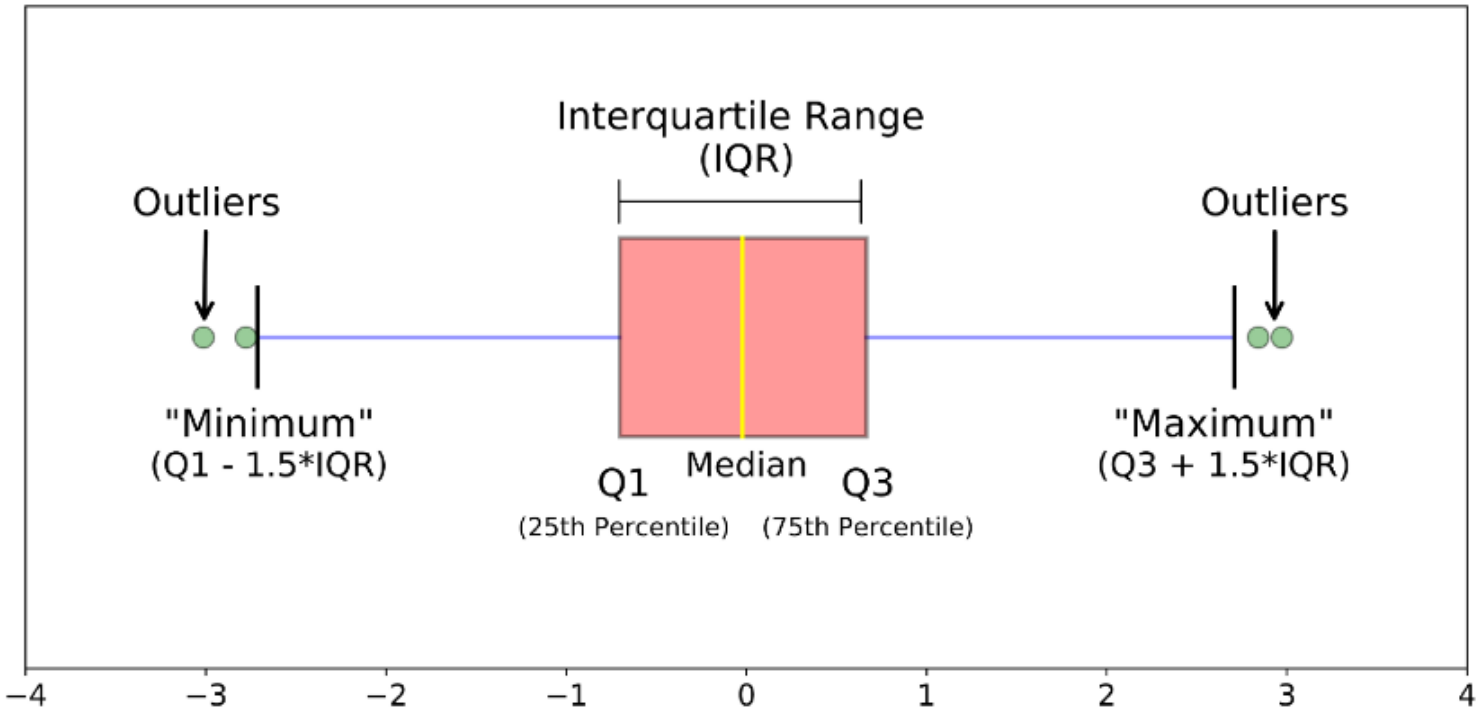
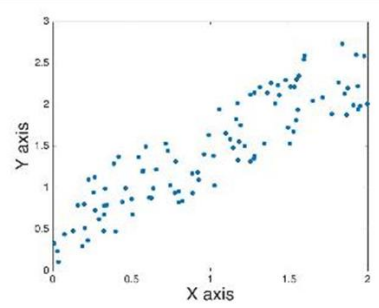
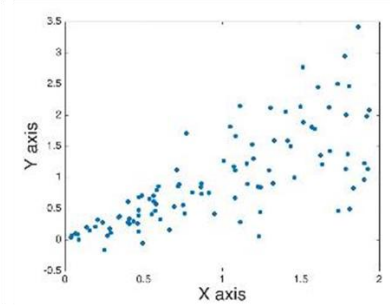


Diagrama de dispersión

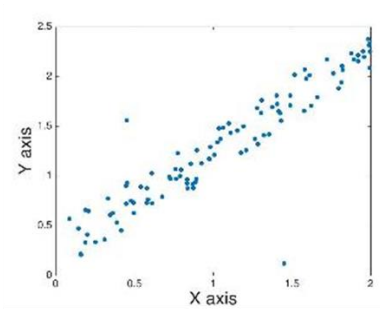
- Diagrama más común para datos bivariados
- Muestra gráficamente:
  - Dependencia entre X e Y
  - Relación lineal o no lineal
  - Si la variación en Y depende de X
  - Datos atípicos



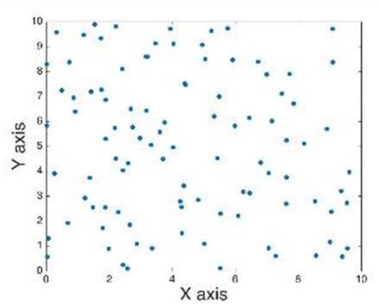
Homocedasticidad



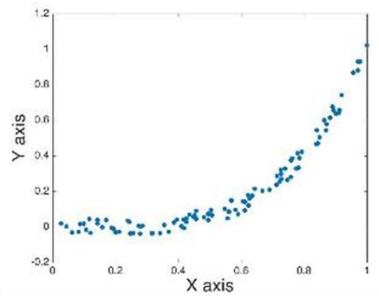
Heterocedasticidad



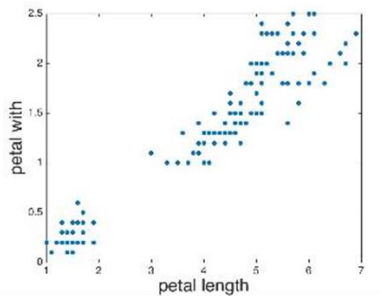
Datos atípicos



Independencia



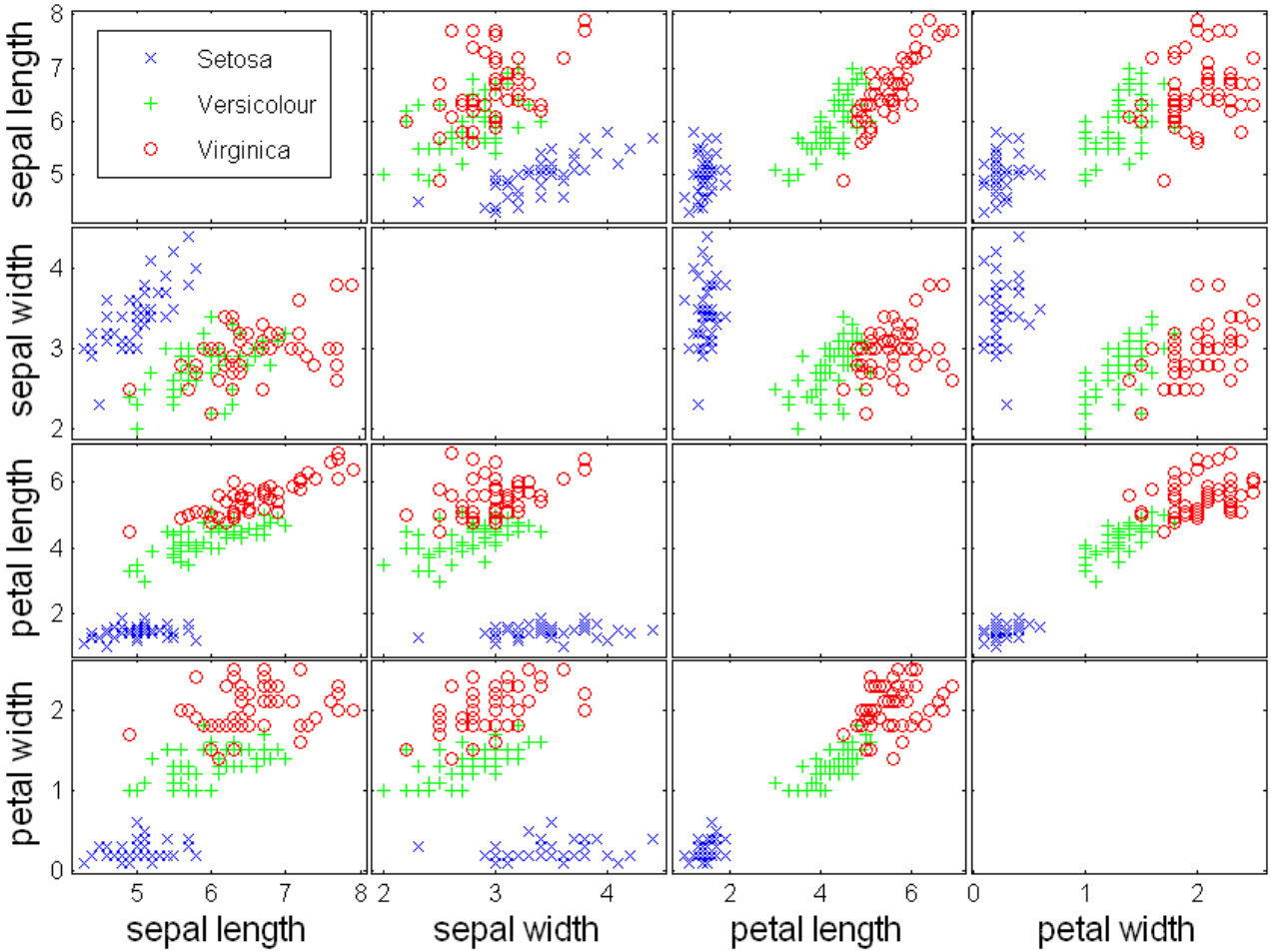
Relación no lineal



Relación lineal

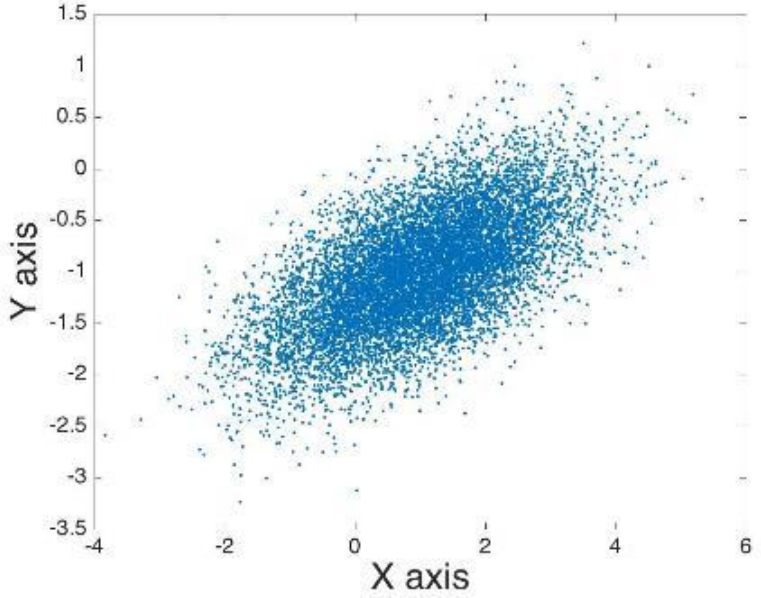
Matriz de dispersión

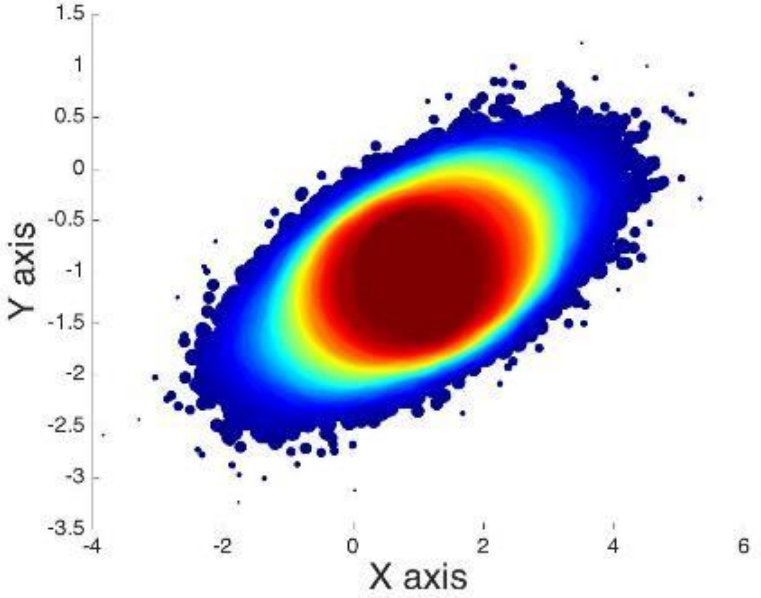
Visualiza todas las combinaciones posibles de variables



Gráficos de densidad bidimensional

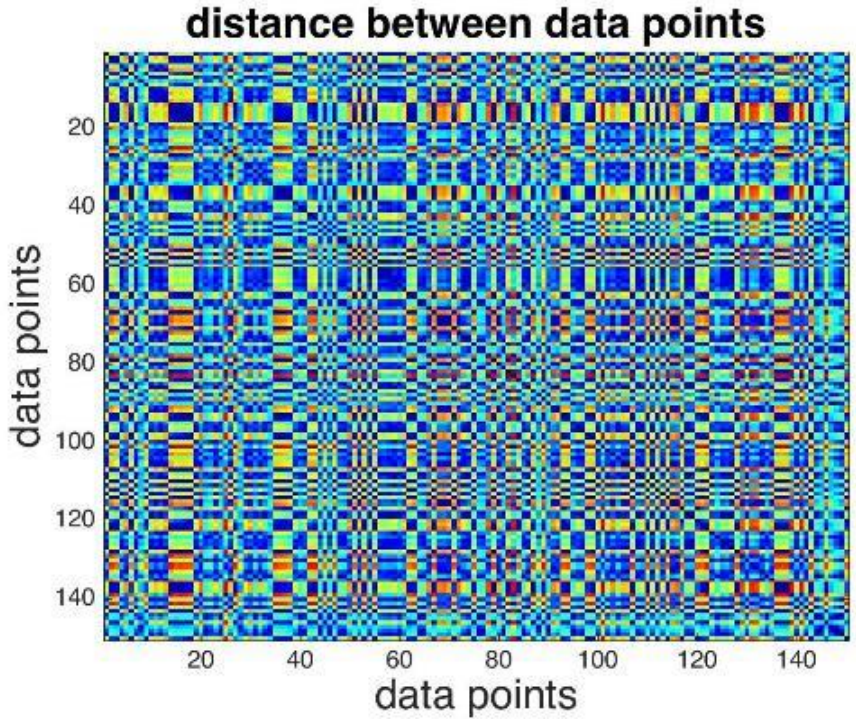
- Utiliza un núcleo bivariado para estimar la densidad bidimensional
- Permite resolver dos limitaciones de los diagramas de dispersión:
  - Visualización de los mismos puntos varias veces
  - Demasiados puntos para discernir la relación.

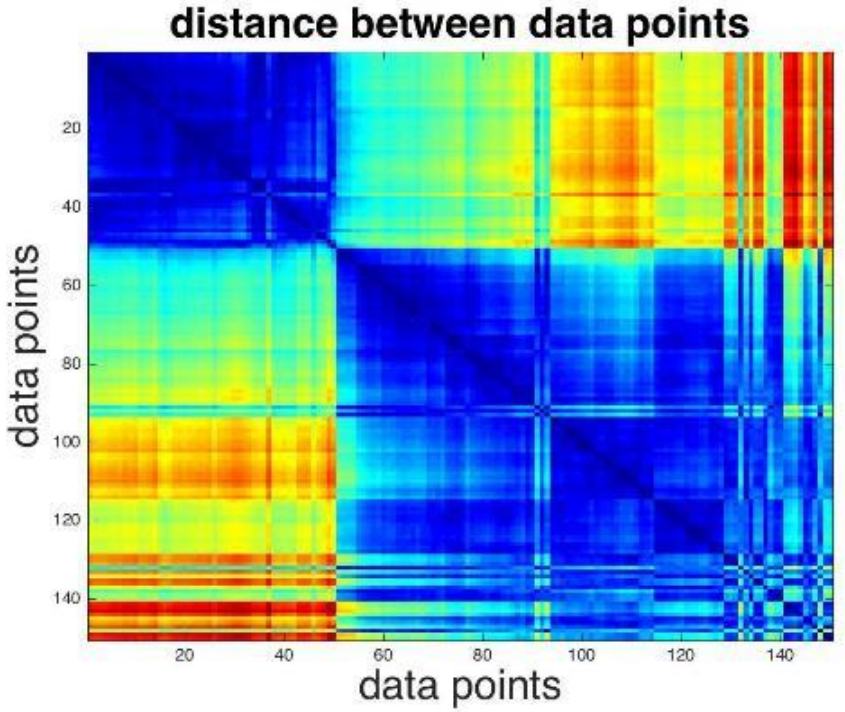


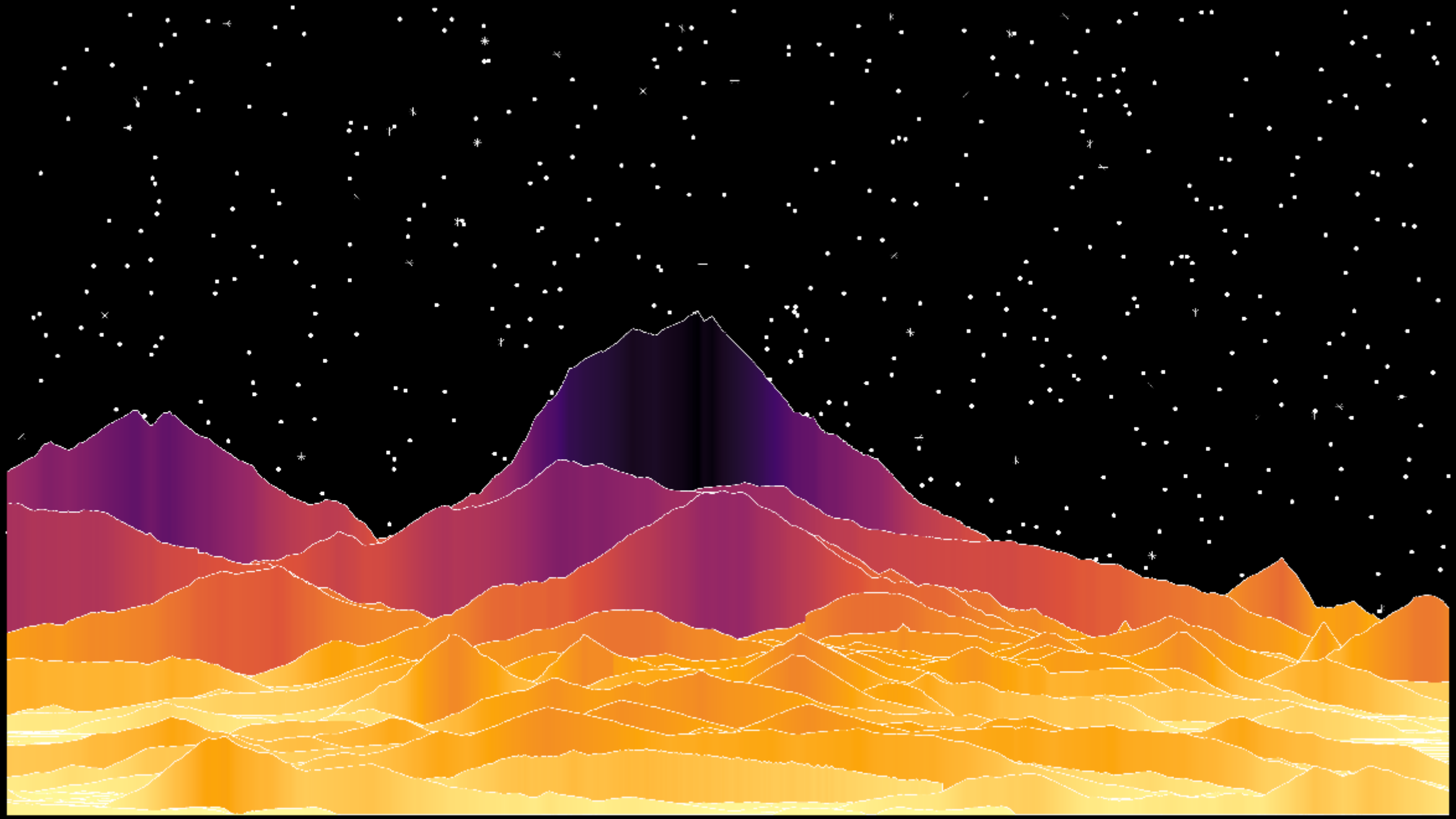


Matriz de distancias

Muestra gráficamente la distancia entre todos los puntos del conjunto de datos







Evitar errores comunes

- Demasiados colores (o colores para daltónicos)
- Demasiada data
- Escalas truncadas
- Textos inadecuados
- Diagramas inadecuados
- Seleccionar solo data favorable (Cherry picking)

Exploración de datos

Dr. Raimundo Sánchez

raimundo.sanchez@uai.cl

@raimun2