Mirando los datos

María Inés Stimolo y Pablo Ortiz

Base de datos

VARIABLE

Observación

	Order	PID	MSSubClass	MSZoning	LotFrontage	LotArea
2673	2674	903225090	50	RM	50.0	5000
303	304	910205120	50	RM	50.0	9140
1478	1479	907418010	20	RL	85.0	11058
2874	2875	910203100	30	RM	61.0	8534
334	335	923251080	20	RL	NaN	26142
894	895	908203090	20	RL	64.0	6410
575	576	533253070	120	RL	61.0	3782
1386	1387	905200100	190	RL	60.0	12900
355	356	527162120	60	RL	60.0	7500
1417	1418	905480240	50	RL	60.0	9084

Variable es una característica que varía en cada observación. Cada valor que asume una variable es un **dato**.

DATO

- De imagen
- De lenguaje natural
- De sensores
- Transaccionales

Base de datos o matriz de datos

VARIABLE: columna

Observación fila

$$\mathbf{X}_{n\times p} = \begin{array}{c|cccc} \text{Obs} & \mathbf{X}_1 & \mathbf{X}_2 & \cdots & \mathbf{X}_p \\ \hline 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 2 & x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \ddots & \cdots \\ n & x_{n1} & x_{n2} & \cdots & x_{np} \end{array} \qquad \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \cdot \\ \mathbf{x}_n' \end{bmatrix}$$

$$(\mathbf{x}_1 \quad \mathbf{x}_2 \quad \ldots\ldots \quad \mathbf{x}_p)$$

Espacio de observaciones R^n

Espacio de variables R^p

DATO
Un elemento de la matriz

María Inés Stimolo y Pablo Ortiz

Tipos de variables y escalas de medición

Tipos de variables

Categóricas o cualitativas

string data is always stored with an object dtype

Numéricas o cuantitativas

Discretas

int_	Default integer type (same as C long; normally either int64 or int32)
intc	Identical to C int (normally int32 or int64)
intp	Integer used for indexing (same as C ssize_t; normally either int32 or int64)
int8	Byte (–128 to 127)
int16	Integer (–32768 to 32767)
int32	Integer (–2147483648 to 2147483647)
int64	Integer (–9223372036854775808 to 9223372036854775807)
uint8	Unsigned integer (0 to 255)
uint16	Unsigned integer (0 to 65535)
uint32	Unsigned integer (0 to 4294967295)
uint64	Unsigned integer (0 to 18446744073709551615)

Continuas

float_	Shorthand for float64
float16	Half-precision float: sign bit, 5 bits exponent, 10 bits mantissa
float32	Single-precision float: sign bit, 8 bits exponent, 23 bits mantissa
float64	Double-precision float: sign bit, 11 bits exponent, 52 bits mantissa
complex_	Shorthand for complex128
complex64	Complex number, represented by two 32-bit floats

Niveles de medición

- Nominal
- Ordinal
- De intervalos o De Razón

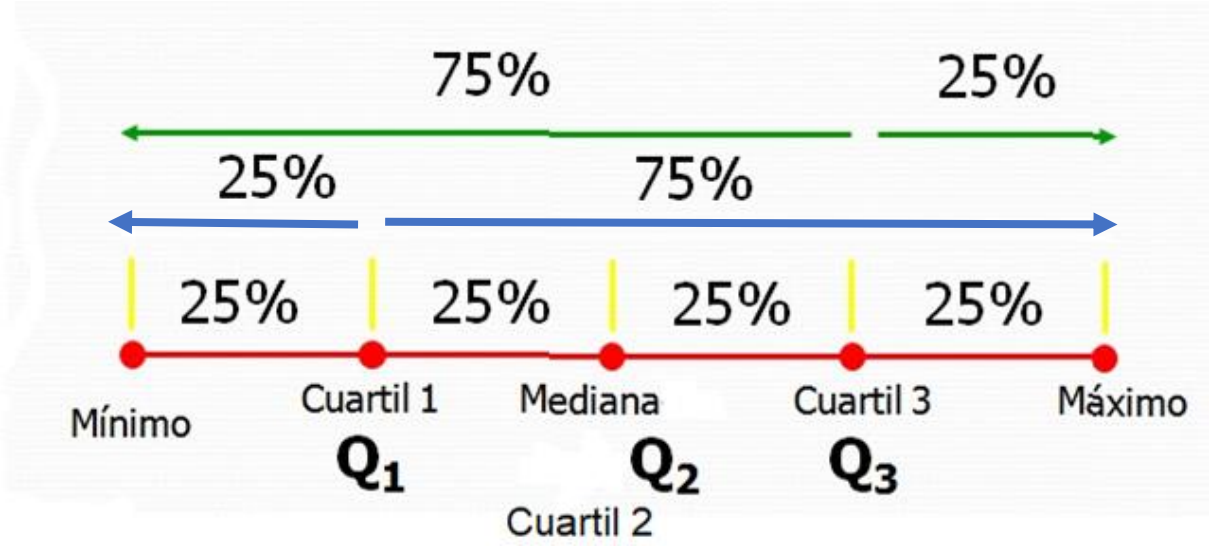
María Inés Stimolo y Pablo Ortiz

¿Que usamos?

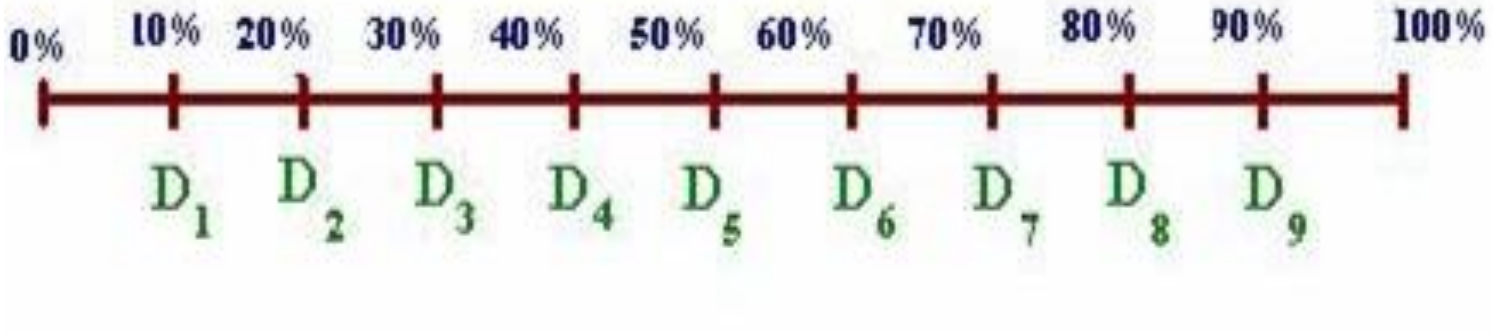
Tipo de variable	Tabla de frecuencias	Gráfico	Medidas descriptivas		
			Posición	Dispersión	Forma
Categórica	si	▪ Barras ▪ Tortas ▪ Diagrama de Pareto	▪ Moda ▪ Mediana (sólo cuando es ordinal)	Desviación Mediana (sólo cuando es ordinal)	
Cuantitativa Discreta	si tiene pocos valores	▪ Bastones ▪ Escalonado ▪ Tallo y hoja ▪ Box Plot	▪ Media ▪ Mediana ▪ Cuartiles ▪ Moda	▪ Varianza ▪ Desviación estándar/ mediana ▪ Coeficiente de variación	▪ Asimetría ▪ Curtosis
Cuantitativa Continua	no	▪ Histograma ▪ Ojiva ▪ Box Plot	▪ Media ▪ Mediana ▪ Cuartiles ▪ Intervalo Modal	▪ Varianza ▪ Desviación estándar/mediana ▪ Coeficiente de variación	▪ Asimetría ▪ Curtosis

Medidas de tendencia no central

Cuartiles

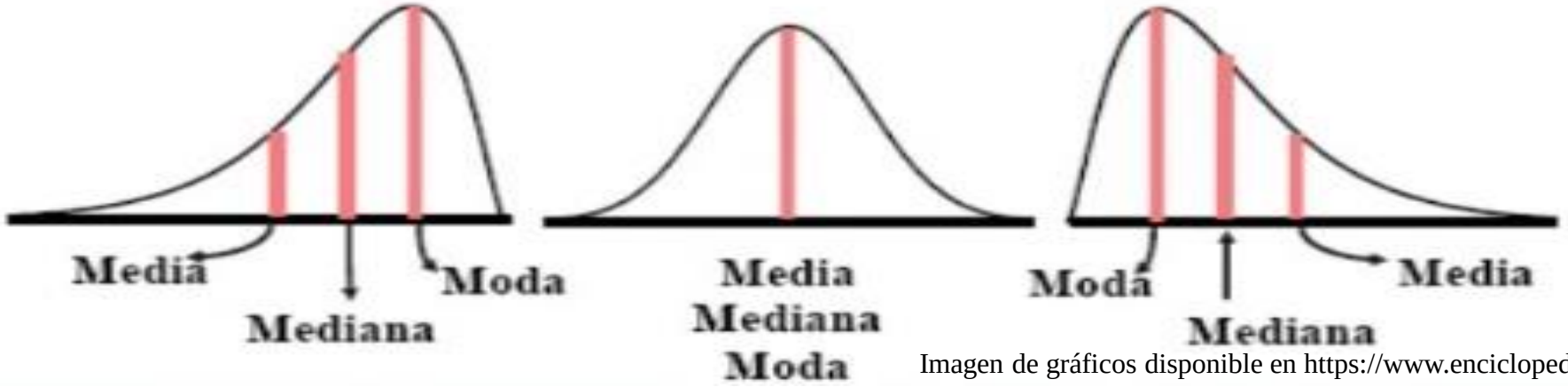


Deciles



Asimetría

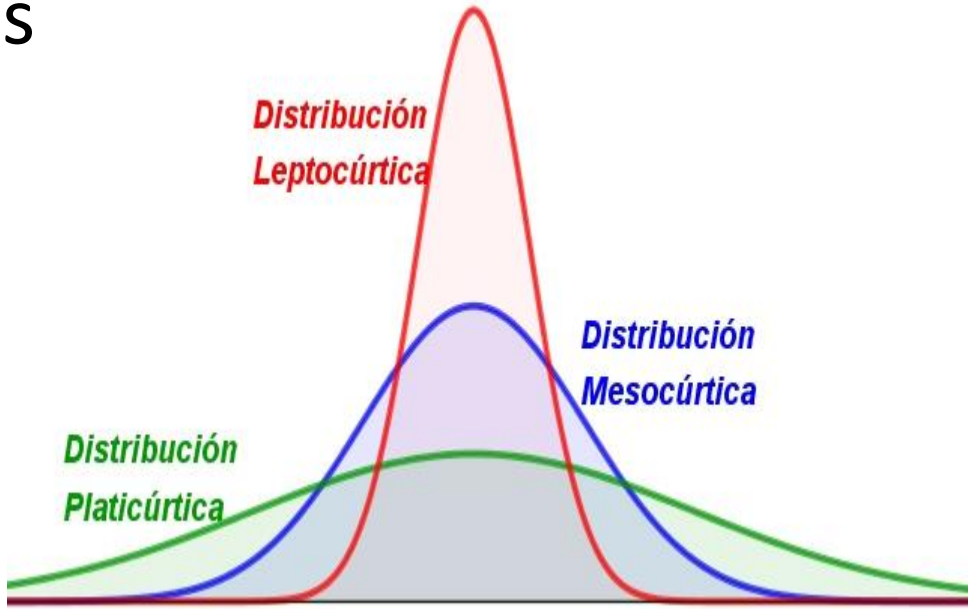
Asimétrica izquierda | Simétrica | Asimétrica derecha



La relación entre medidas de posición cambia según la asimetría

Imagen de gráficos disponible en https://www.enciclopediafinanciera.com/images/asimetria.jpg

Curtosis

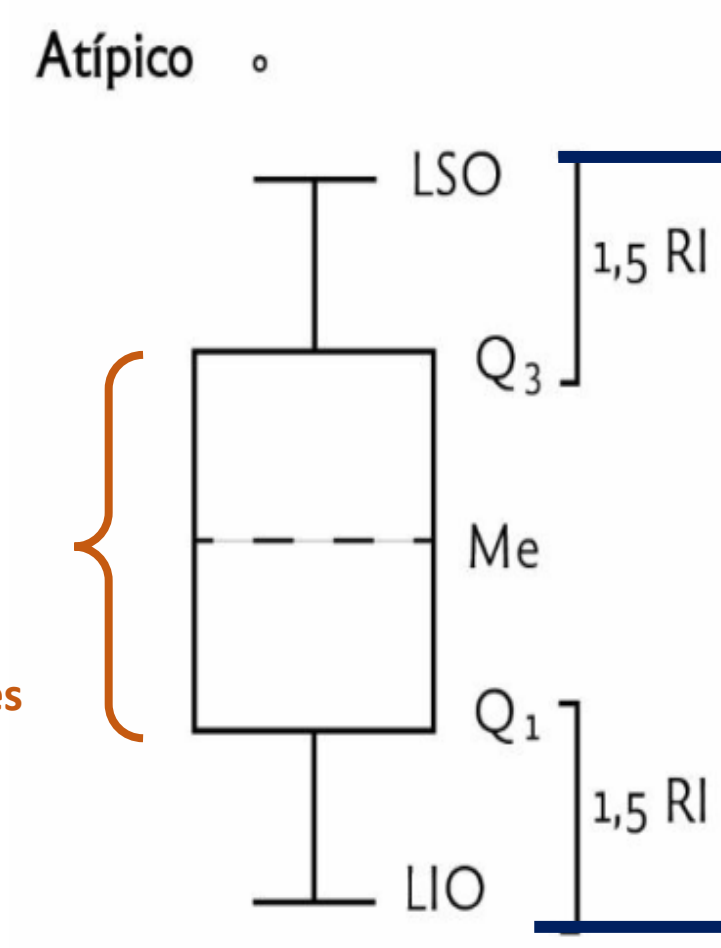


Imagendisponible en https://www.lifeder.com/wp-content/uploads/2020/03/curtosis.jpg

Gráfico de caja y brazos (box plot)

Permite ver la forma de la distribución y detectar casos atípicos

Rango Intercuartílico → $RI = Q_3 - Q_1$



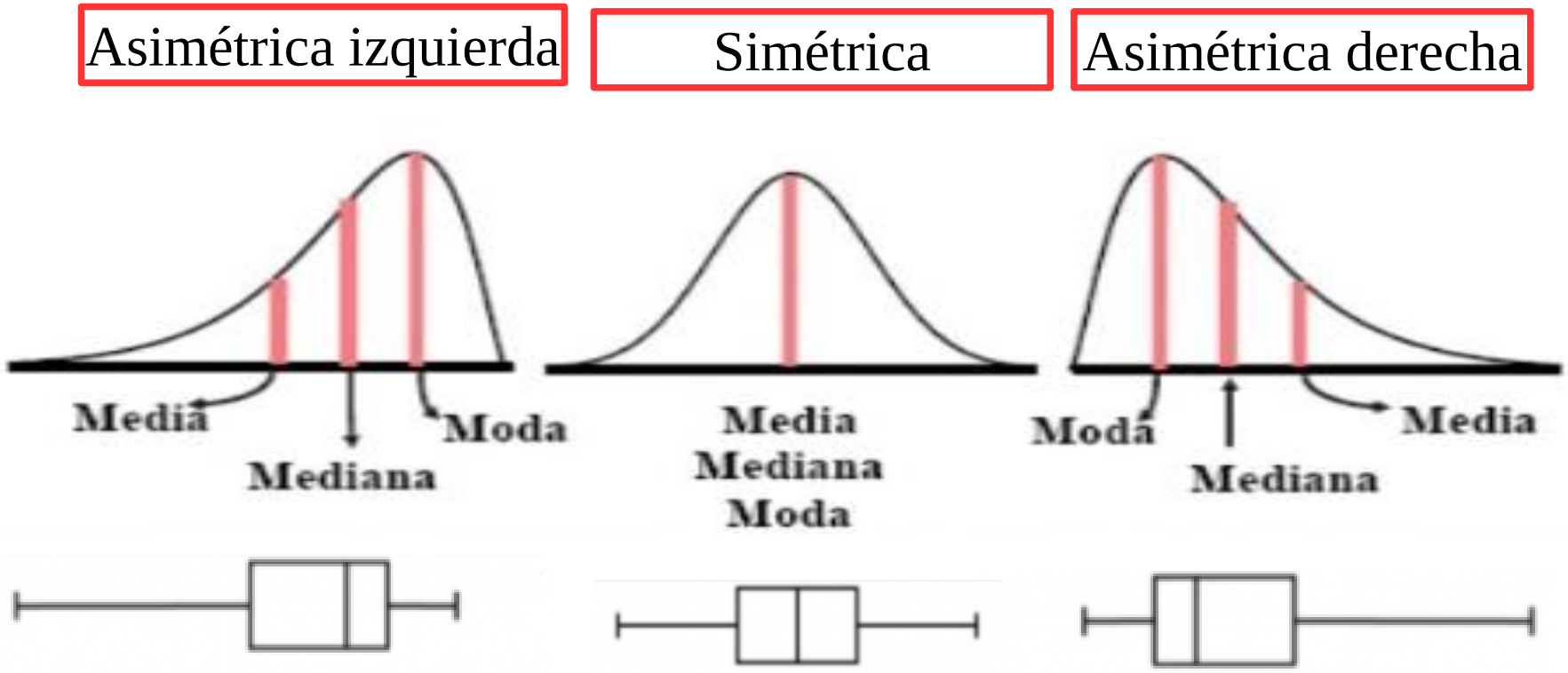
Límite superior atípicos: Q_3 +1,5 RI (teórico)

Estos limites son teóricos, el brazo llega al valor de la última observación anterior a estos límites.
Atípicos: toda observación que este fuera de estos límites

Caja: para observar la distribución del 50% central de las observaciones

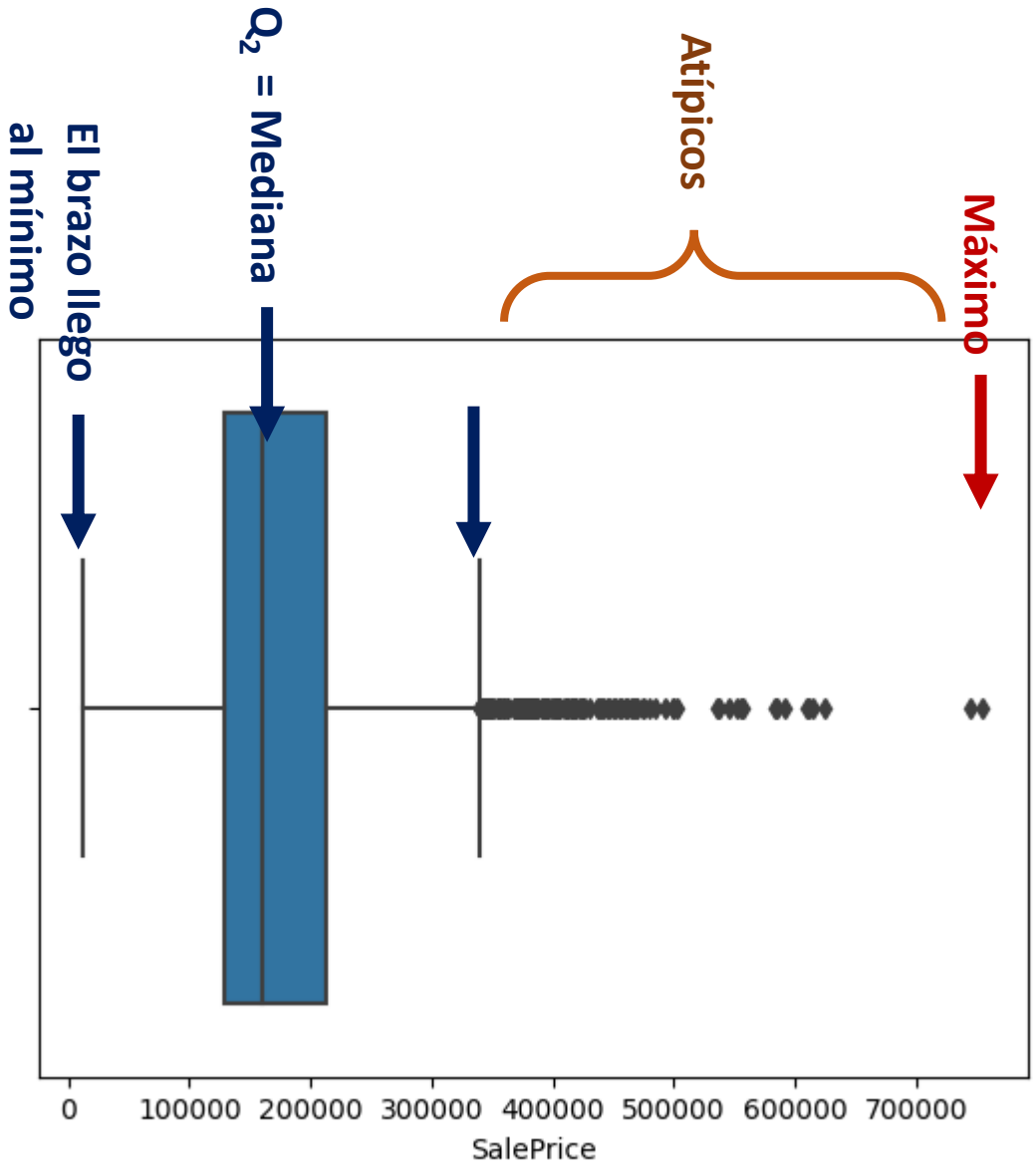
Límite inferior atípicos: Q_1 - 1,5 RI (teórico)

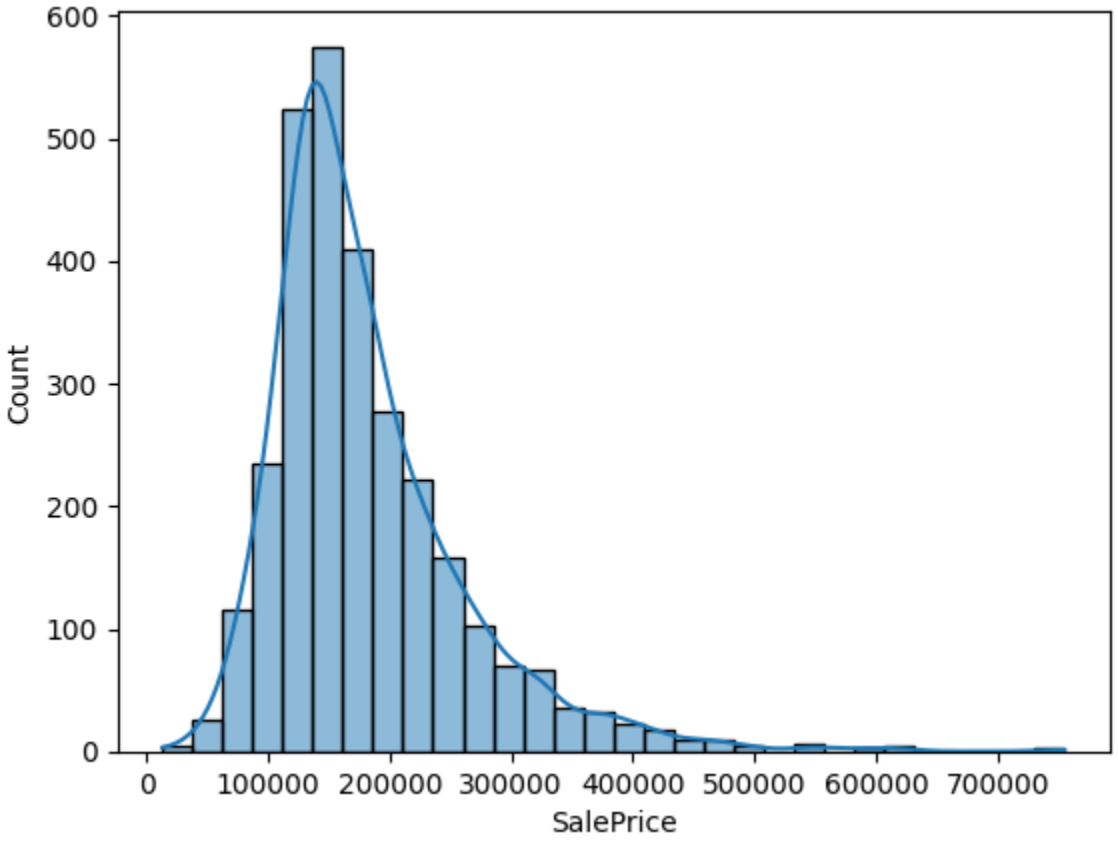
Gráfico de caja y brazos (box plot)



María Inés Stimolo y Pablo Ortiz

Gráfico de caja y brazos (box plot)





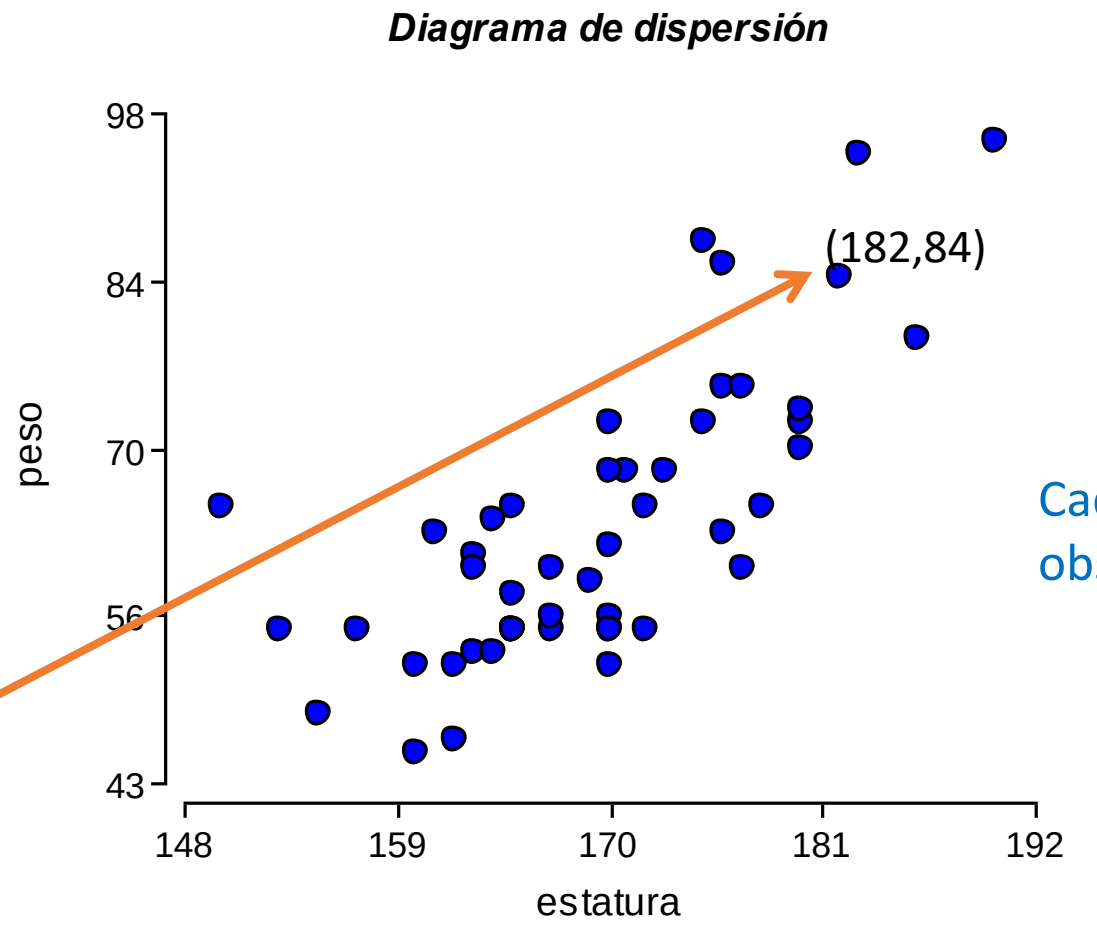
Mirando el comportamiento de variables en forma conjunta

María Inés Stimolo y Pablo Ortiz

Dos variables numéricas. Gráfico de dispersión

Permite ver la relación conjunta de las dos variables y detectar datos atípicos

peso	estatura
56	170
55	157
65	150
68	171
72	175
53	163
55	165
61	163
65	165
48	155
62	170
72	180
46	162
84	182
85	176
70	180



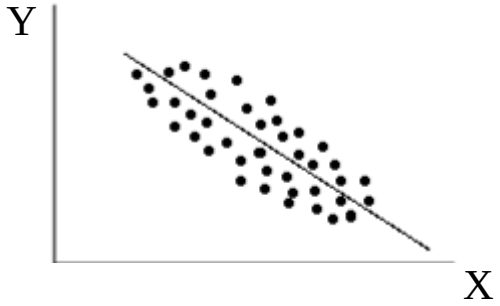
Cada punto es una observación

María Inés Stimolo y Pablo Ortiz

Bivariadas: Medidas de Asociación lineal entre variables numéricas

Covarianza Cov(x,y) – Coeficiente de correlación $R_{x,y}$

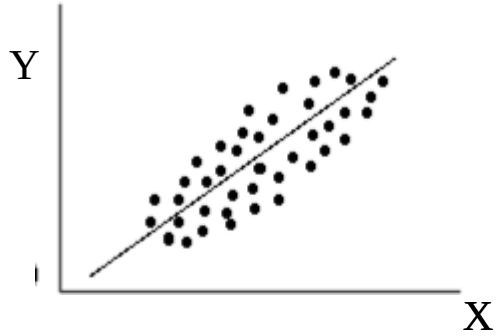
Asociación negativa



$$-\infty < Cov(x,y) < 0$$

$$-1 \leq R_{x,y} < 0$$

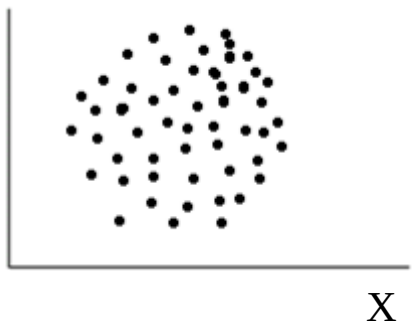
Asociación positiva



$$0 < Cov(x,y) < \infty$$

$$0 < R_{x,y} \leq 1$$

No hay asociación



$$Cov(x,y) = 0$$

Relación no lineal

$$R_{x,y} = 0$$

Matriz de varianzas-covarianzas. Resumen de variabilidad entre *p* variables

$$S = \frac{1}{n-1}\tilde{X}'\tilde{X}$$

$\tilde{X}$ es una matriz de orden $n \times p$ donde cada una de las p columnas es una variable restada su media aritmética (variable en desvío)

$\tilde{X}'\tilde{X}$ es la matriz de sumas de cuadrados y producto cruzados y representa la variabilidad de los datos (contiene la variabilidad de cada variable y la variabilidad conjunta entre pares de variables)

$$S = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \cdots & \cdots & \ddots & \cdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix}$$

Vector de desvíos $\mathbf{d}_j = \mathbf{x}_j - \bar{x}_j \mathbf{1}_n$ $\mathbf{d}_j = \begin{bmatrix} x_{1j} - \bar{\mathbf{x}}_j \\ x_{2j} - \bar{\mathbf{x}}_j \\ . \\ x_{nj} - \bar{\mathbf{x}}_j \end{bmatrix}$

Varianza $s_j^2 = s_{jj} = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}{n-1} = \frac{\sum_{i=1}^{n} d_{ij}^2}{n-1} = \frac{\mathbf{d}_j' \mathbf{d}_j}{n-1}$

Covarianza $s_{jk} = s_{kj} = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{n-1} = \frac{\mathbf{d}_j' \mathbf{d}_k}{n-1}$ $j,\ k = 1,\ 2,.., p$

María Inés Stimolo y Pablo Ortiz

Matriz de correlación

Es la matriz de varianzas y covarianzas de los datos estandarizados $\frac{X-\mu_X}{\sigma_X}$ donde μ_X es la media aritmética y σ_X el desvío estándar de la variable X.

$$\mathrm{R} = \mathrm{D}^{-\frac{1}{2}}\mathrm{S}\mathrm{D}^{-\frac{1}{2}} = \begin{bmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & 1 & \dots & r_{2p} \\ \dots & \dots & \ddots & \dots \\ r_{p1} & r_{p2} & \dots & 1 \end{bmatrix}$$

Donde $\mathrm{D} = diag(s_{11}, s_{22} \dots\dots s_{pp})$

Coeficiente de correlación

$$r_{jk} = r_{kj} = \frac{s_{jk}}{\sqrt{s_{jj}}\sqrt{s_{kk}}}$$

Datos perdidos o faltantes ¿qué hacemos?

Si los métodos de análisis se basan en información completa, sin datos perdidos, tenemos algunas alternativas

Eliminarlos Algunos algoritmos utilizan este mecanismo por defecto. En este caso, se debe analizar si los faltantes son aleatorios. Esto es, asegurarse que no se están eliminando casos con características particulares, lo que derivaría en un sesgo relevante al momento del análisis y que la cantidad de observaciones retenidas sea suficiente para que el algoritmo funcione correctamente.

Imputarle un valor asignar a los datos faltantes valores obtenidos bajo algún criterio adecuado. Se pueden definir distintos criterios:

Sustitución de casos*:* En los casos en que se pueda buscar una nueva observación, cada observación con datos faltantes es reemplazada por una nueva no incluida inicialmente y para la cual se dispongan de datos completos.
Valores representativos: Consiste en asignar a cada dato faltante un valor relativo a los datos observados, como puede ser el promedio, la mediana, el mínimo, el máximo o, incluso, un valor aleatorio dentro del rango para cada variable.
Utilizar otras observaciones completas: se reemplazan los datos faltantes por valores calculados a partir de uno o más observaciones completas del mismo conjunto de datos. Existen diferentes formas de asignar eligiendo el valor de una observación al azar o calculando la media de valores correspondientes de un grupo de observaciones.
Crear un modelo predictivo para estimar valores que sustituirán los datos faltantes. La variable con datos faltantes se usa como variable respuesta o predicha, y las variables restantes se usan como entrada para el modelo predictivo de clasificación o regresión, según el tipo de variable con datos faltantes.

Cómo detectamos datos atípicos

El **diagrama de caja y brazos** permite detectar atípicos en forma univariada, el **diagrama de dispersión** en forma bivariada. Se han definido muchos métodos de detección de datos atípicos analizando las variables en forma conjunta. Incluso muchos de ellos tienen que ver con el problema que se plantea y el algoritmo utilizado.

PyOD: Librería Python para Detección de Outliers disponible en https://pyod.readthedocs.io/en/latest/ basado en el trabajo de Songqiao Han y otros (https://www.andrew.cmu.edu/user/yuezhao2/papers/22-neurips-adbench.pdf) dan un listado de funciones para detectar datos atípicos.

¿qué hacemos una vez que los identificamos?

Eliminarlos si el objetivo del análisis es tener un resultado sobre datos considerados típicos o normales.

Analizar la sensibilidad probando el algoritmo con y sin los casos atípicos. Esto permite analizar el efecto de los atípicos en los resultados.

Analizarlos por separado en caso de que se pueda identificar que el conjunto de datos atípicos corresponden a un grupo con características bien diferenciadas.