

Tema 2: Representación gráfica de variables

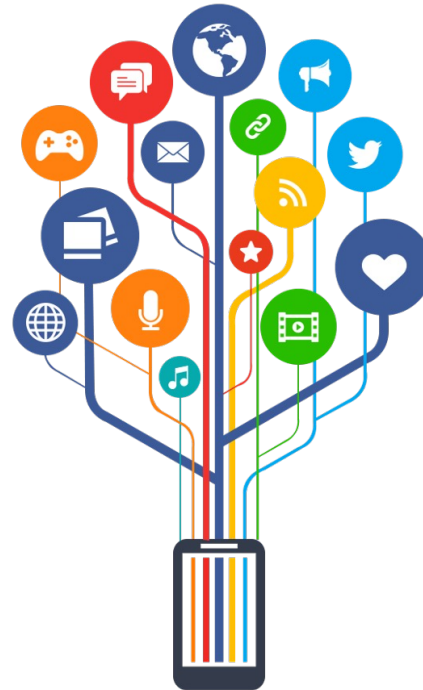
Minería de Datos

1. Proceso de KDD. Etapas. MD vs otras disciplinas . Tipos de conocimiento. Ejemplos.
2. Análisis de datos. Tipos de variables. Descripciones estadísticas. Representaciones gráficas. Ejemplos.

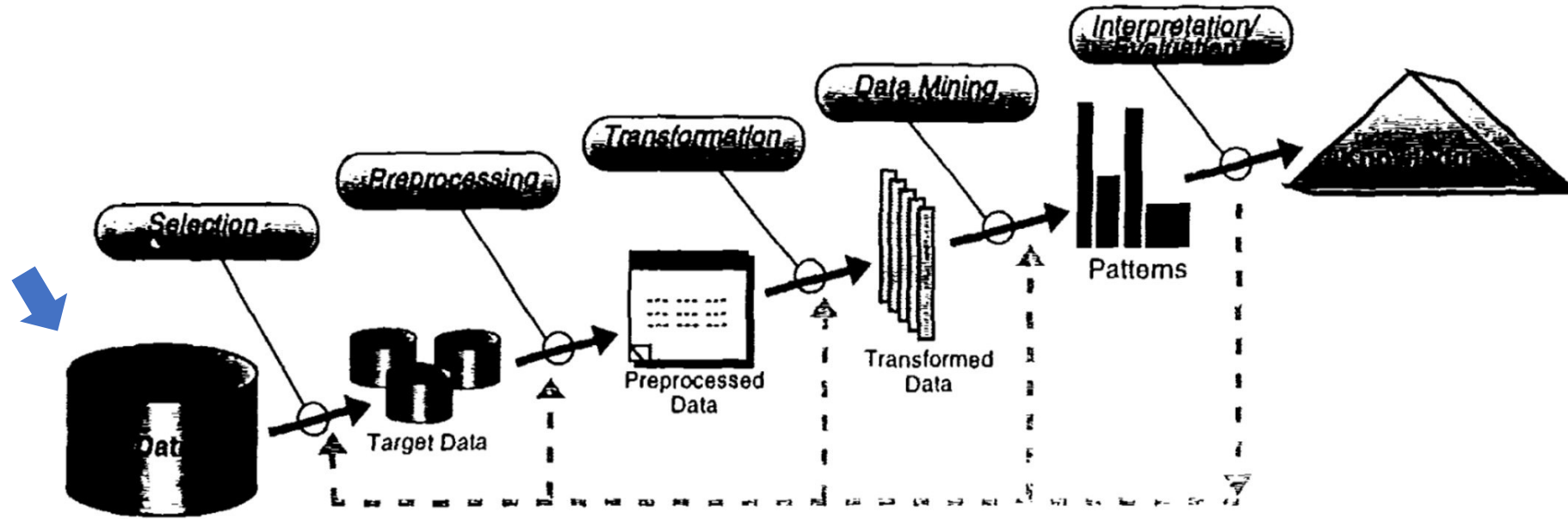
1. Proceso de KDD. Etapas. MD vs otras disciplinas . Tipos de conocimiento. Ejemplos.

2. Análisis de datos. Tipos de variables. Descripciones estadísticas. Representaciones gráficas. Ejemplos.

> ¿Cómo se originan los datos?



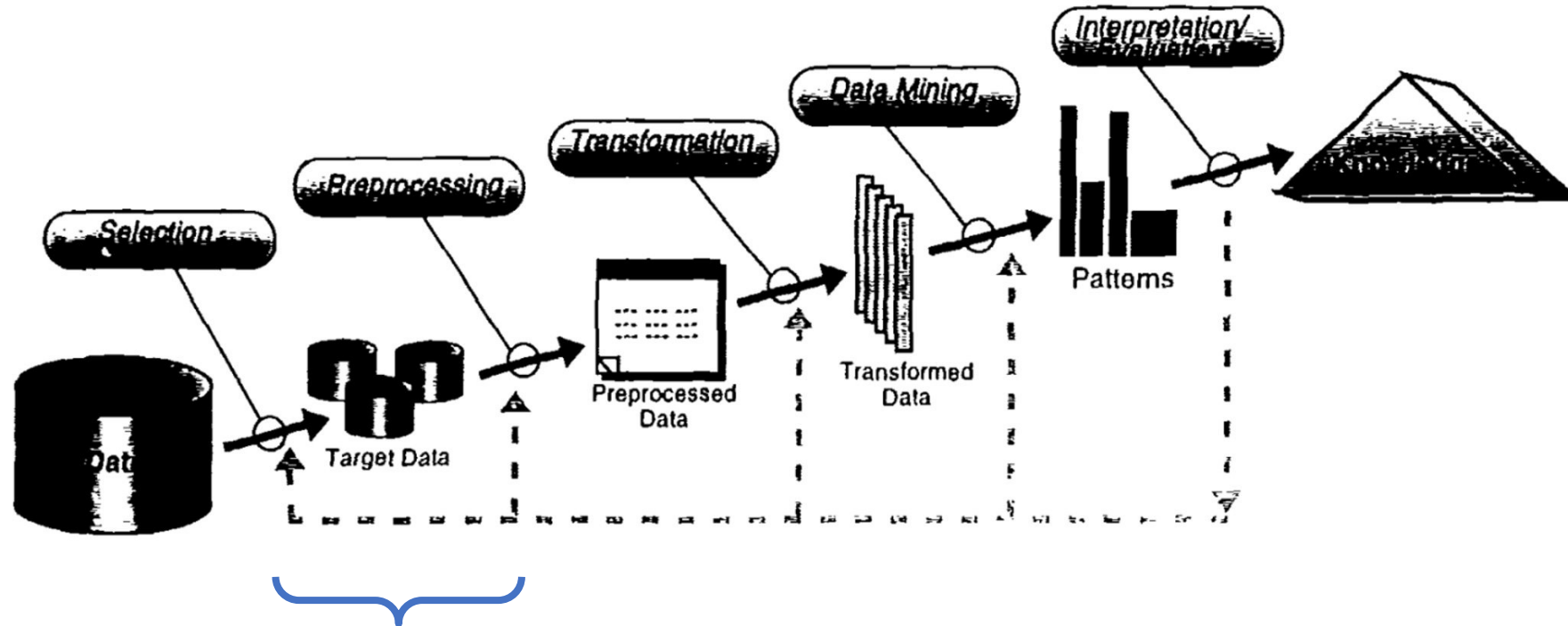
> Etapas del proceso de KDD (según Fayaad)



Generalmente son datos registrados en forma previa al proceso de KDD

- Almacena información histórica
- No necesariamente centralizada
- Variedad de BBDD y formatos

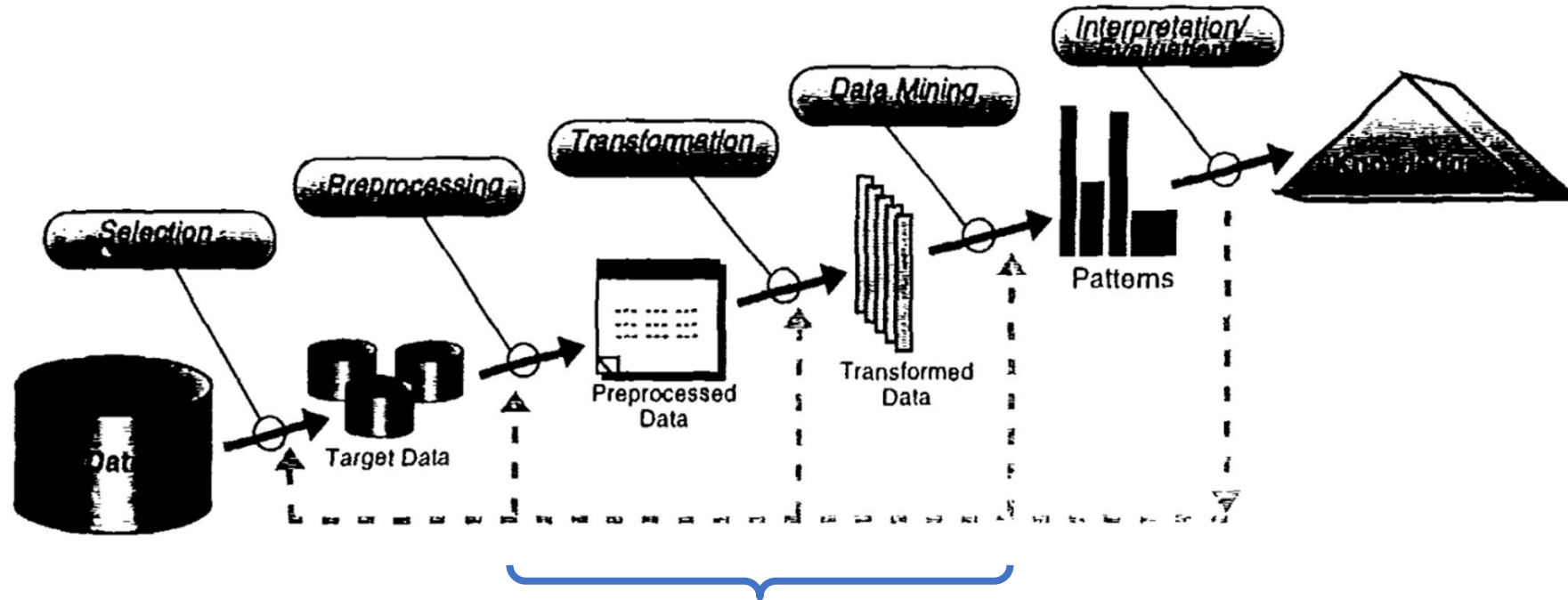
> Etapas del proceso de KDD (según Fayaad)



Datos seleccionados:

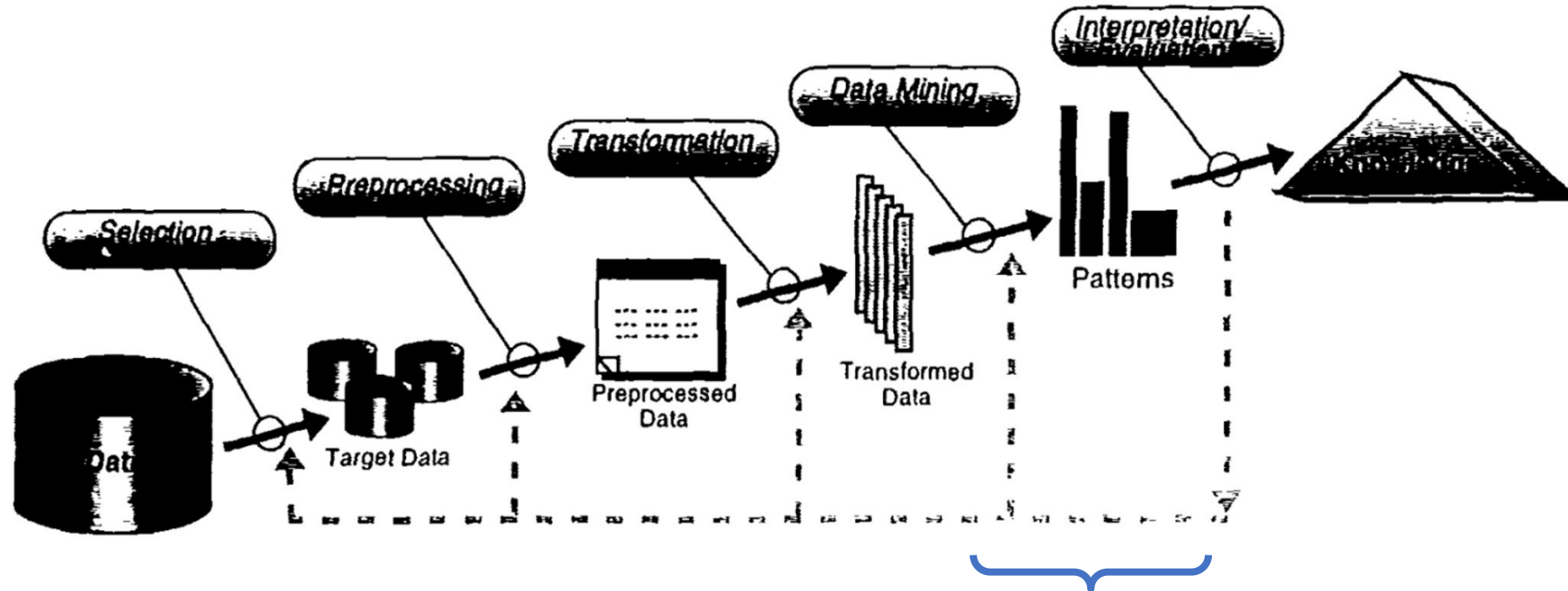
- Elegidos en base al problema
- Medidas subjetivas y objetivas

> Etapas del proceso de KDD (según Fayaad)



- Uniformar la notación
- Datos faltantes
- Fuera de los rangos esperados (outliers)
- Nuevos atributos

> Etapas del proceso de KDD (según Fayaad)



ÁRBOL



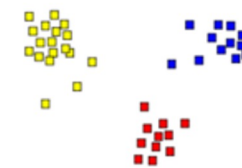
REGLAS

IF (TIPO = CC) AND (SODIO > 470)
ENTONCES (COSTO=BAJO)

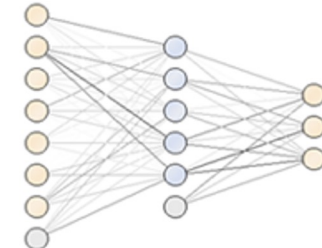
IF (TIPO = CR) AND (PRODUCTO = CN)
ENTONCES (COSTO=ALTO)

IF (TIPO = DC)
ENTONCES (COSTO=MEDIO)

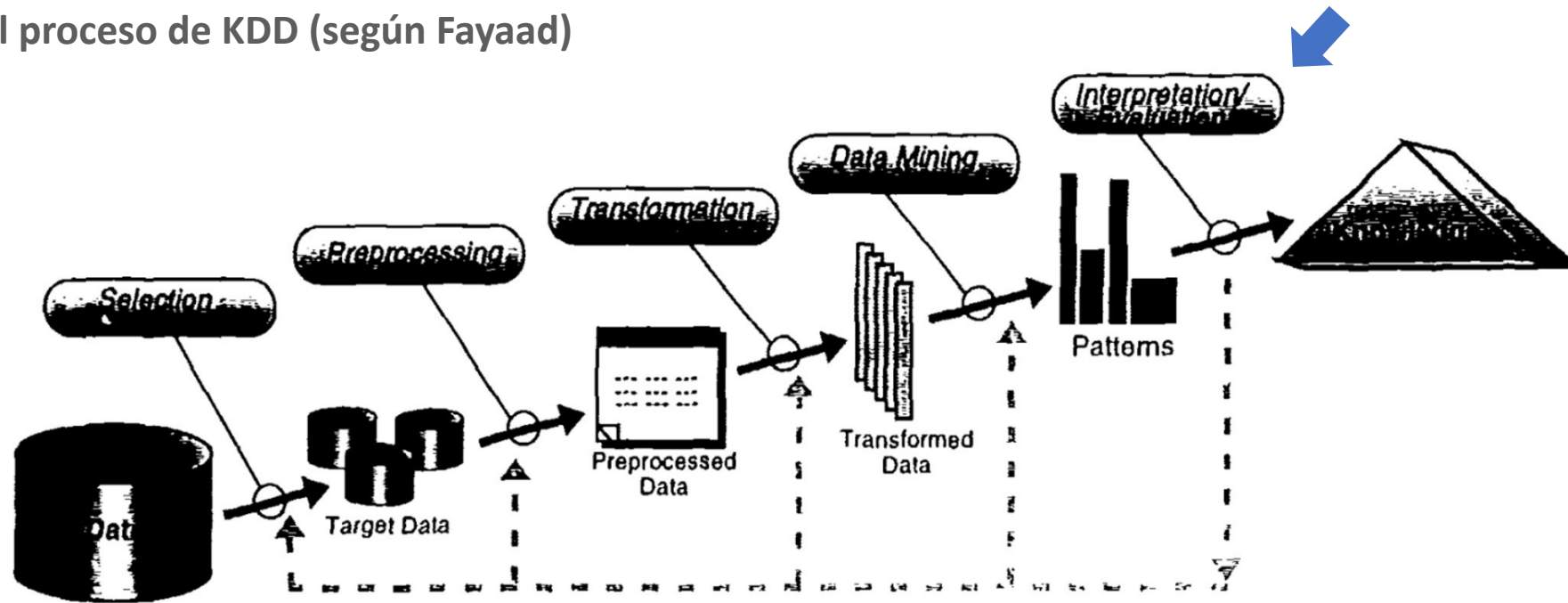
AGRUPAMIENTO



RED NEURONAL

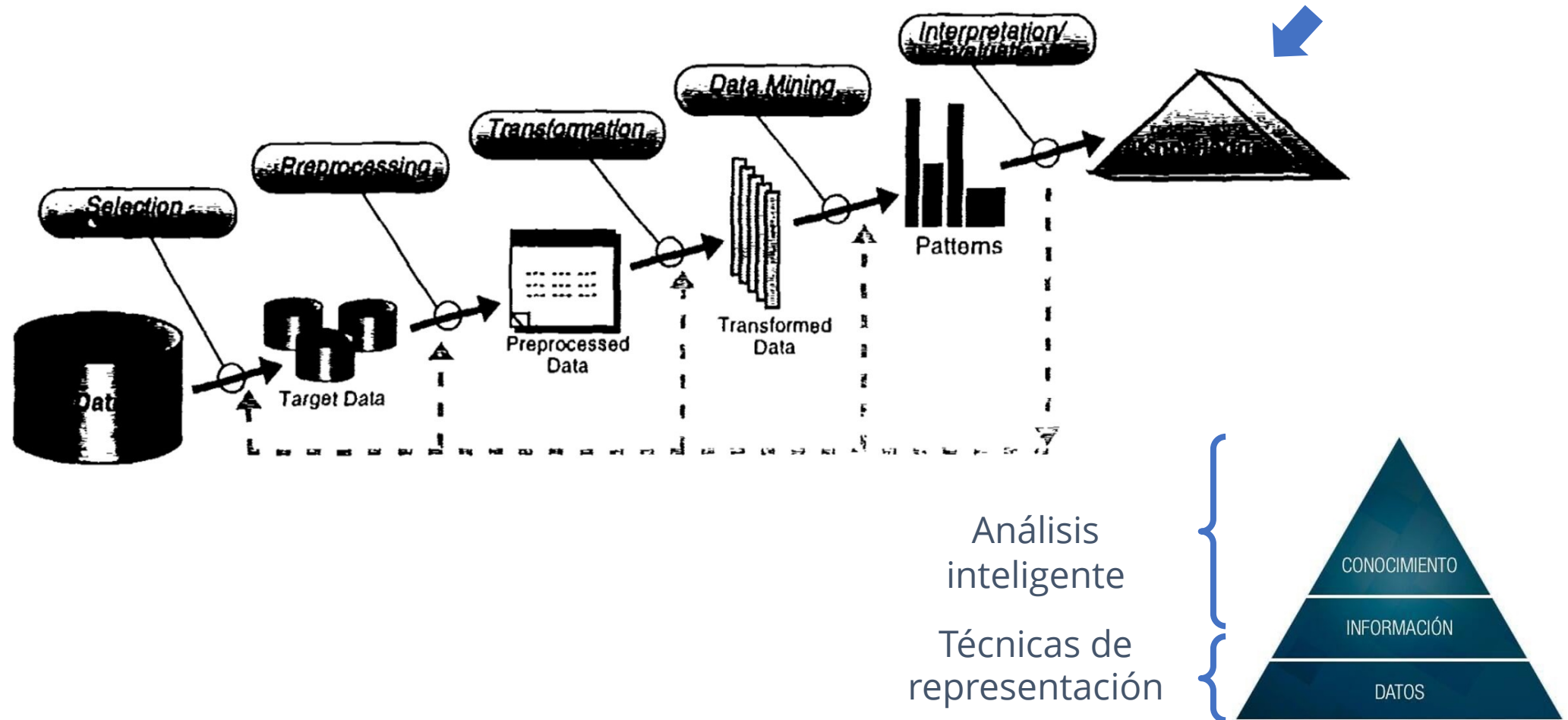


> Etapas del proceso de KDD (según Fayaad)



- El objetivo de la validación es **medir el desempeño** del modelo
- Se busca **determinar la calidad** de la respuesta brindada
- Para que la medición sea objetiva debe hacerse sobre un conjunto de datos diferente al utilizado para generar el modelo

> Etapas del proceso de KDD (según Fayaad)



> Ejemplo de datos y modelo

Variables
(columnas / atributos / características /
dimensiones)

Registros
(filas / ejemplos /
observaciones / instancias)

ASISTENCIA	TRABAJA	INGRESO	FORO	RESULTADO
15	0	DESAP	NO	DESAP
15	0	DESAP	SI	DESAP
20	0	APROB	NO	APROB
5	0	APROB	SI	APROB
20	23	DESAP	NO	DESAP
10	10	DESAP	SI	DESAP
0	50	APROB	NO	APROB
12	40	APROB	SI	APROB
65	0	DESAP	NO	DESAP
75	0	DESAP	SI	APROB
60	30	APROB	NO	APROB
55	40	APROB	SI	APROB
100	15	DESAP	NO	DESAP
80	15	DESAP	SI	APROB
75	20	APROB	NO	APROB
78	12	APROB	SI	APROB

Resultados de
alumnos de un curso

si (INGRESO = APROB) **entonces** (RESULT=APROB)

si (INGRESO = DESAP) **y**
(FORO = NO) **entonces** (RESULT=DESAP)

> Tipos de conocimiento a extraer

Predictivo

- Predice hechos futuros basándose en las variables
- Por ejemplo, se busca predecir:
 - Cuál medicamento suministrarle a un paciente
 - Si un correo electrónico recibido es spam o no
 - El sentimiento de un mensaje

Descriptivo

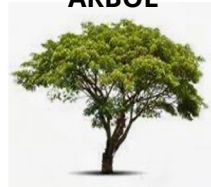
- Muestran nuevas relaciones entre las variables
- Por ejemplo se busca describir:
 - Tipos de clientes para diseñar campañas de marketing
 - Operaciones con tarjeta de crédito para detectar cuáles son fraude

> Tarea predictiva

GATO



ÁRBOL



CUADERNO



**Aprendizaje
supervisado**

GATO



GATO



ÁRBOL



CUADERNO



CUADERNO

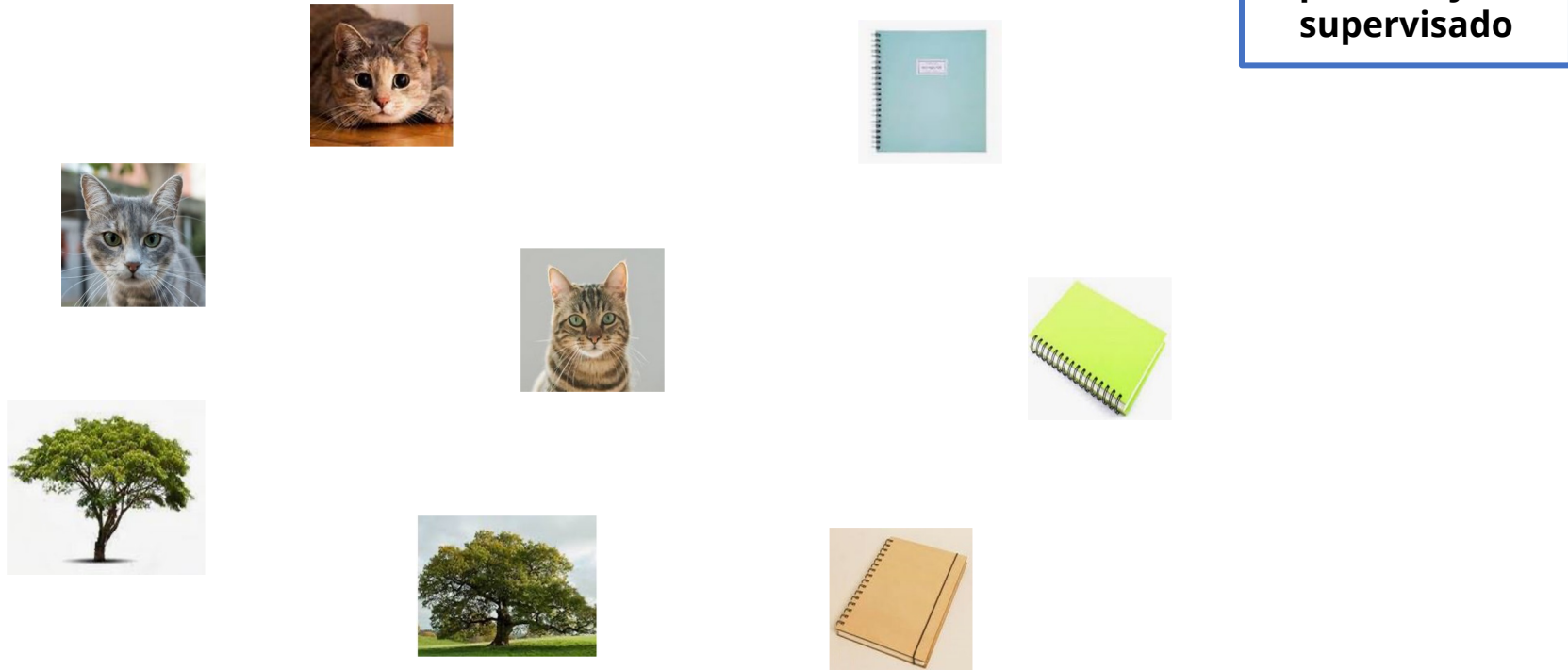


GATO



¿?

> Tarea descriptiva



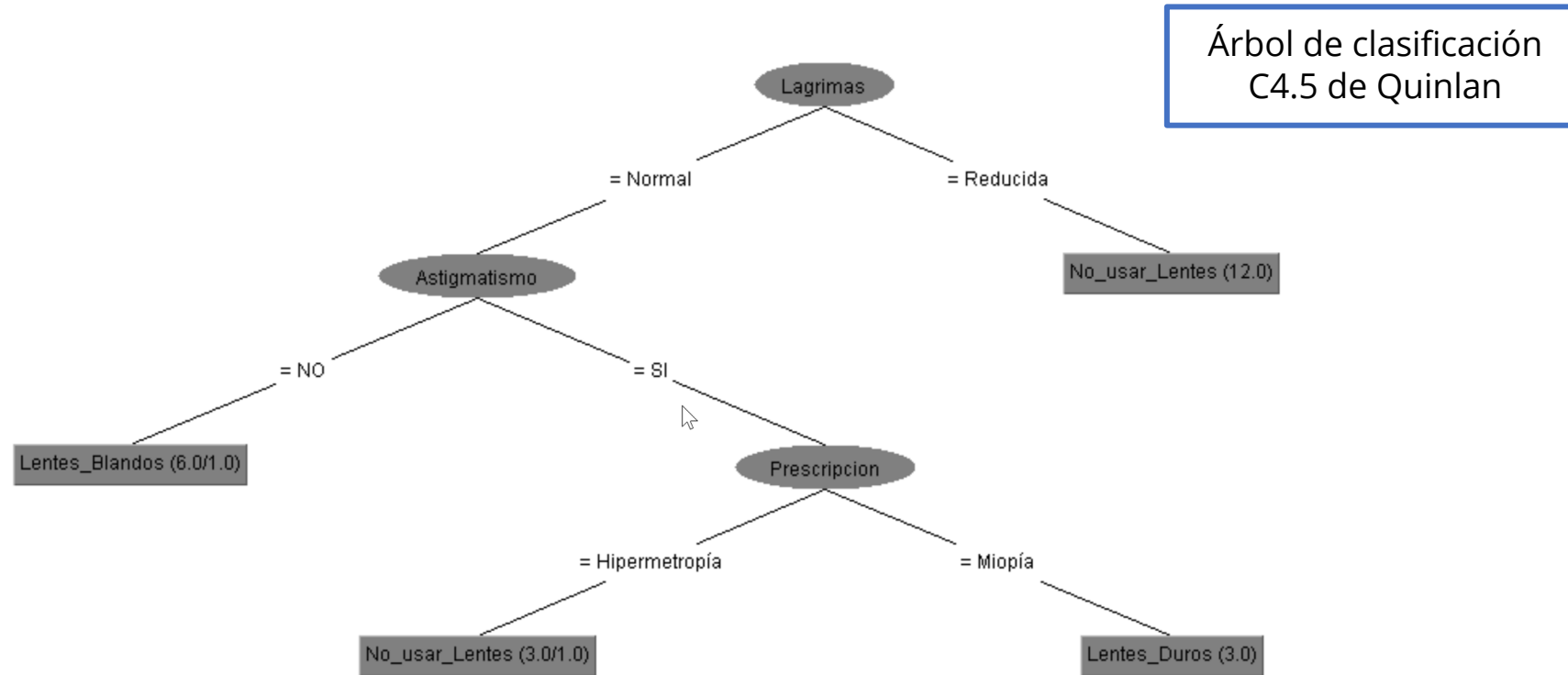


Ejemplo de tarea predictiva: prescripción de lentillas

Fuente: <https://archive.ics.uci.edu/ml/datasets/Lenses>



> Ejemplo de tarea predictiva: prescripción de lentillas

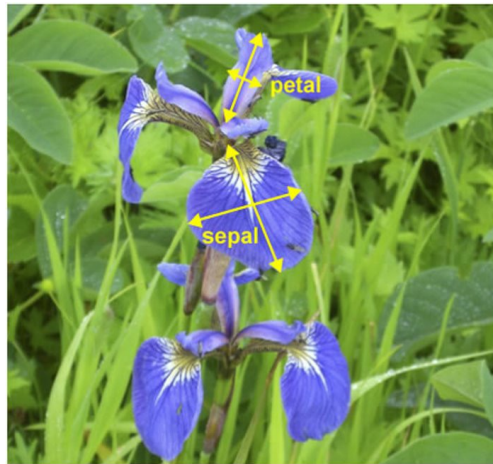




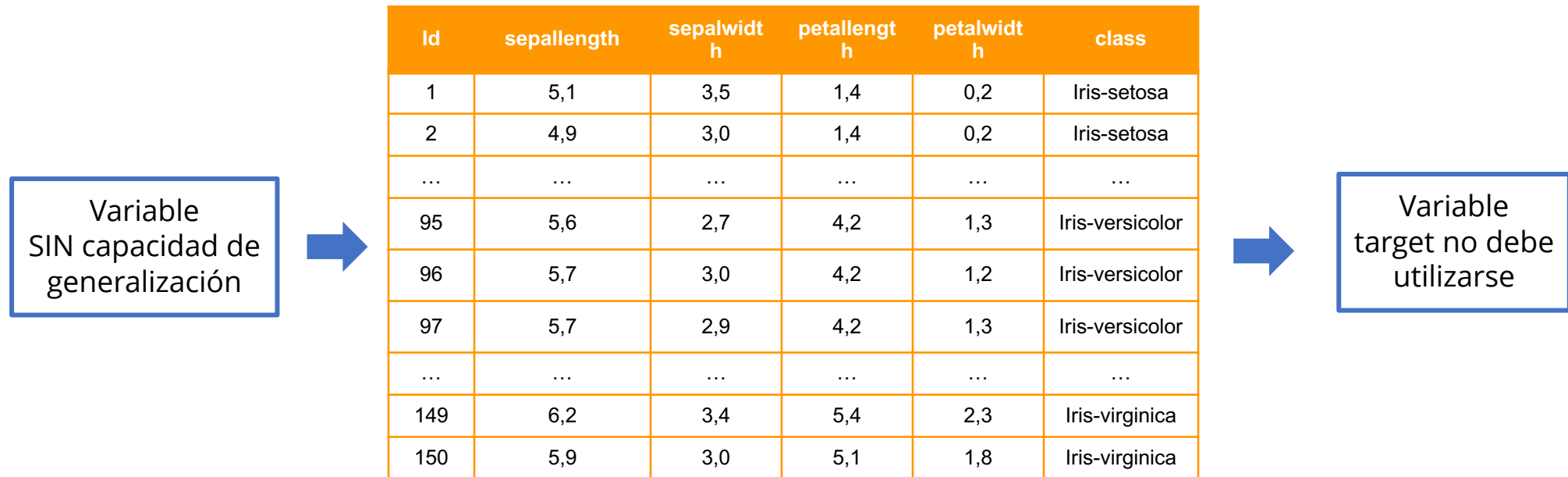
Ejemplo de tarea descriptiva: caracterización de flores

Fuente: <https://archive.ics.uci.edu/dataset/53/iris>

Se dispone de información 3 tipos de flores Iris

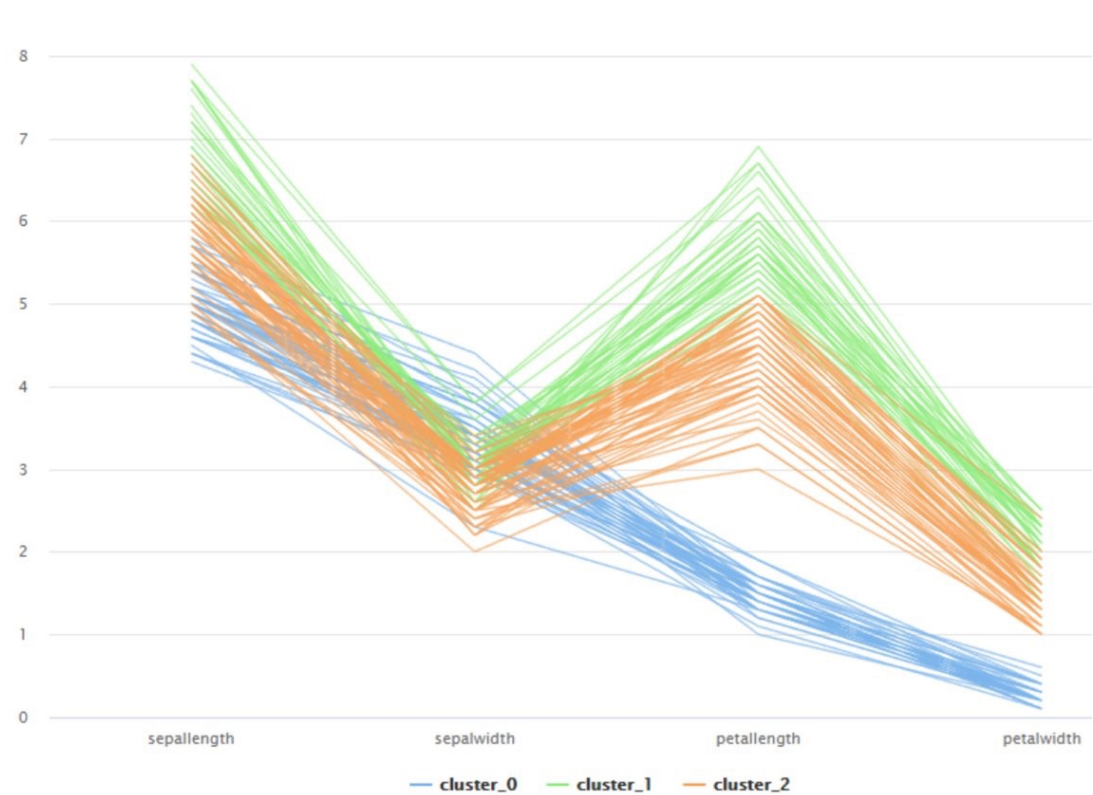


> Ejemplo de tarea descriptiva: caracterización de flores



> Ejemplo de tarea descriptiva: caracterización de flores

Resultado de agrupar los registros:



id	class	cluster	sepal length	sepal width	petal length	petal width
1	Iris-setosa	cluster_0	5.100	3.500	1.400	0.200
2	Iris-setosa	cluster_0	4.900	3	1.400	0.200
3	Iris-setosa	cluster_0	4.700	3.200	1.300	0.200
4	Iris-setosa	cluster_0	4.600	3.100	1.500	0.200

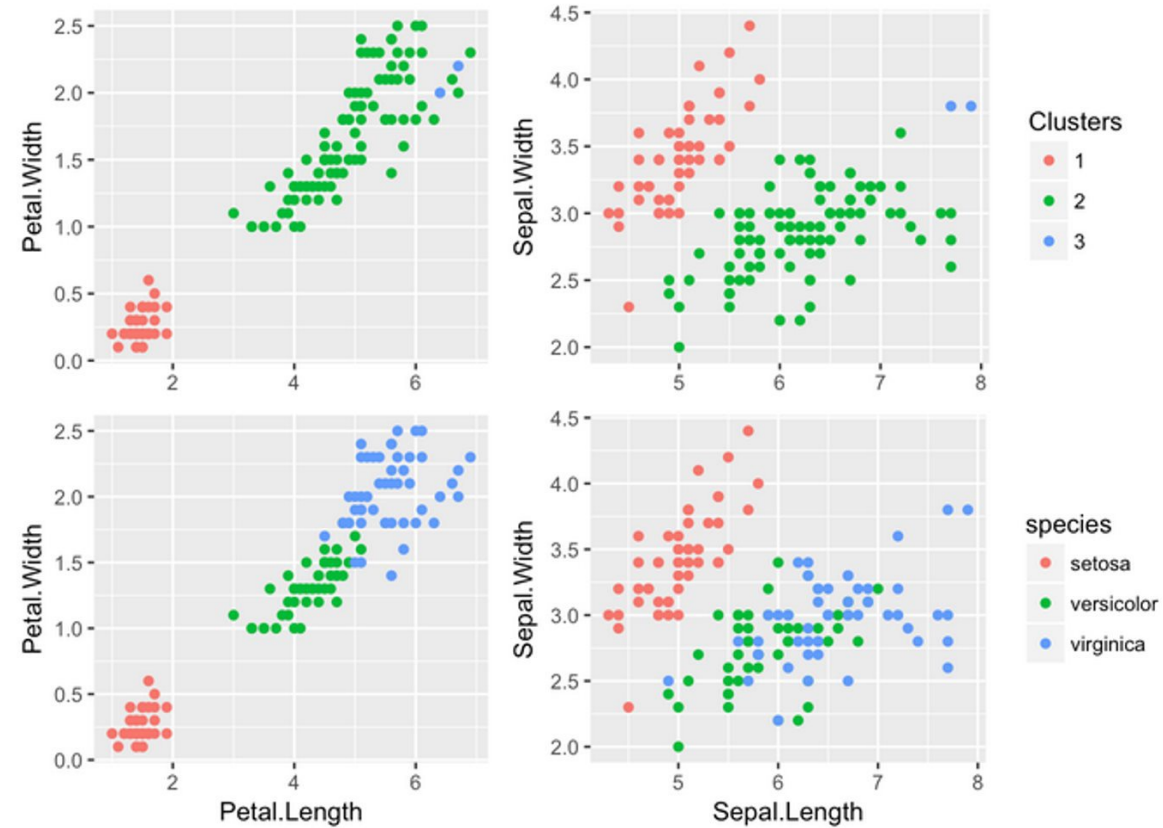
Tabla de centroides

Attribute	cluster_0	cluster_1	cluster_2
sepal length	5.006	6.854	5.884
sepal width	3.418	3.077	2.741
petal length	1.464	5.715	4.389
petal width	0.244	2.054	1.434



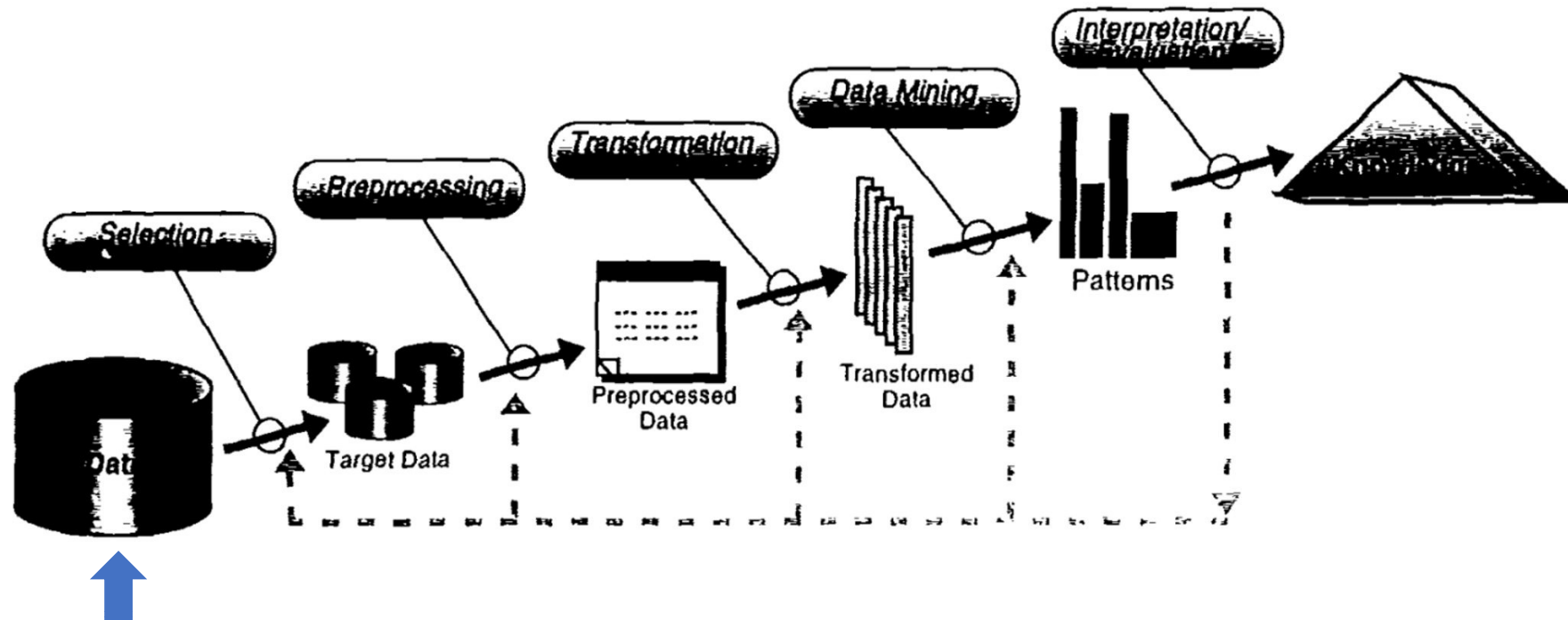
Identificar valores distintos en los diferentes grupos

> Ejemplo de tarea descriptiva: caracterización de flores



1. Proceso de KDD. Etapas. MD vs otras disciplinas . Tipos de conocimiento. Ejemplos.

**2. Análisis de datos. Tipos de variables. Descripciones estadísticas.
Representaciones gráficas. Ejemplos.**



Se comienza analizando los datos disponibles:

- Tipos de variables o atributos
- Medidas y gráficos para conocer su calidad

Cuantitativas o **numéricas**:

- **DISCRETAS**: sin parte decimal (cantidad de empleados, número de visitas, etc.)
- **CONTINUAS**: con parte decimal (sueldo, metros cuadrados, saldo, etc.)

Cualitativas o **categóricas**:

- **NOMINALES**: sin orden entre sus valores posibles (estado civil, idioma, etc.)
- **ORDINALES**: con orden entre sus valores (alto, medio, bajo, etc.)



Ejemplo: datos del consumo de combustible de ciertos autos en ciudad

Fuente: <https://archive.ics.uci.edu/dataset/9/auto+mpg>

- **mpg** : cantidad de millas que puede realizar con un galón de combustible.
- **cylinders**: cantidad de cilindros
- **displacement**: cilindradas
- **horsepower**: potencia del motor

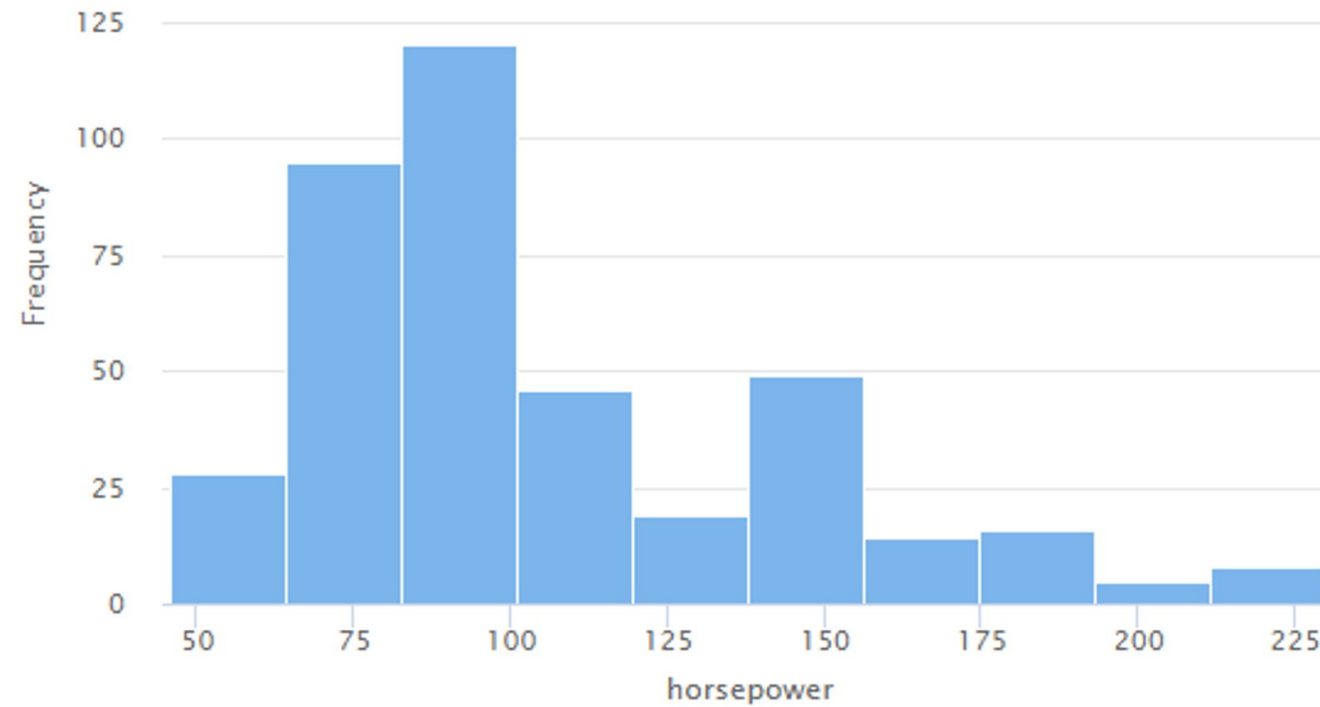
- **weight**: Peso
- **acceleration**: aceleración
- **model_year**: año del modelo
- **origin**: país de fabricación (1-USA, 2-Europe, 3-Japan)
- **car_name**: marca del auto

> Ejemplo: datos del consumo de combustible de ciertos autos en ciudad

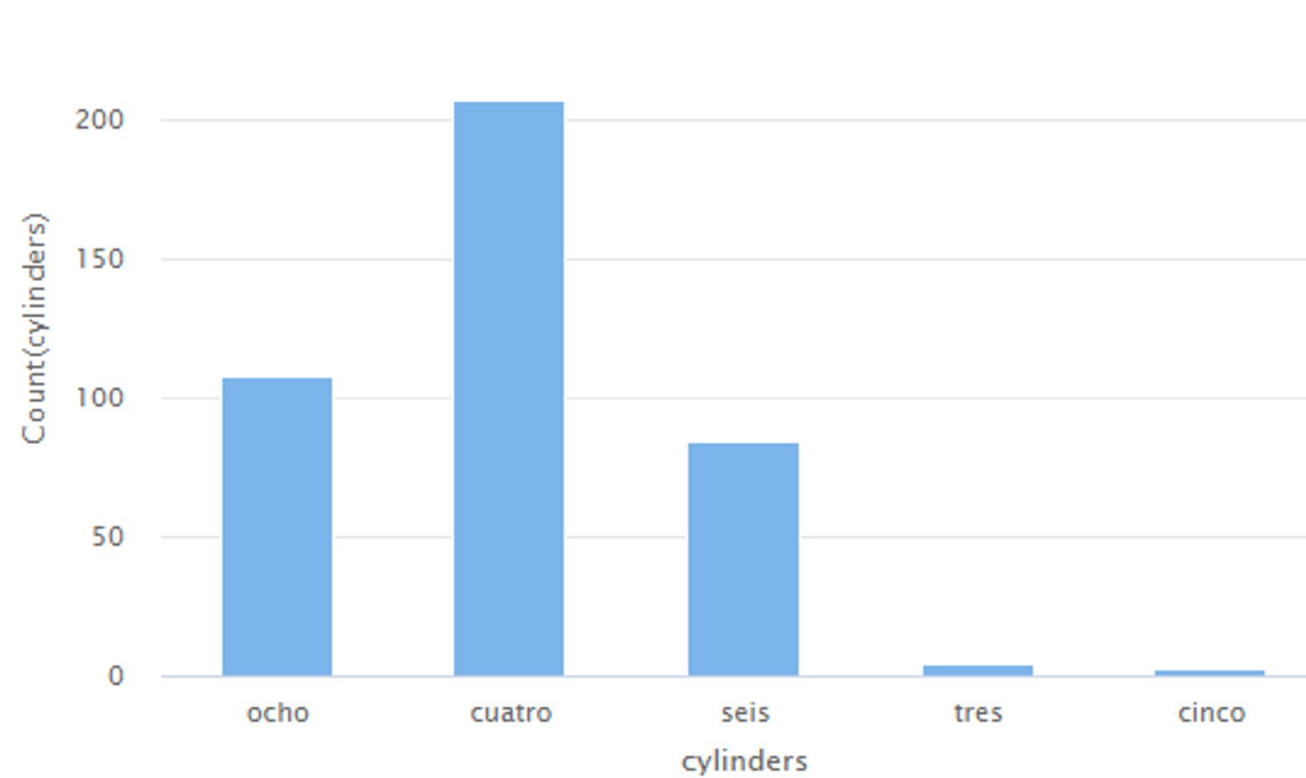
mpg	cylinders	displacement	horsepower	weight	acceleration	model_year	origin	car_name
14.0	ocho	4550	225.0	3086	100	70	1	buick estate wagon (sw)
24.0	cuatro	1130	95.00	2372	150	70	3	toyota corona mark ii
22.0	seis	1980	95.00	2833	155	70	1	plymouth duster
18.0	seis	1990	97.00	2774	155	70	1	amc hornet
...
...
27.0	cuatro	9700	88.00	2130	145	70	3	datson pl510
26.0	cuatro	9700	46.00	1835	205	70	2	volkswagen 1131 deluxe sedan
25.0	cuatro	1100	87.00	2672	175	70	2	peugeot 504

- ¿Cuántos atributos tiene la tabla?
- ¿De qué tipo es cada uno de ellos?

> Histograma de variable numérica: horsepower



> Diagrama de barras de variable categórica: horsepower



Identifican propiedades de los datos y destacan qué valores deben tratarse como ruido o valores atípicos

MEDIDAS DE TENDENCIA CENTRAL

- Media
- Mediana
- Moda
- Rango medio

MEDIDAS DE DISPERSION

- Varianza
- Desviación estándar
- Rango
- Cuartiles
- Rango Inter cuartil

> Media

- La **MEDIA** es el promedio de los valores del atributo. Dicho atributo debe ser numérico.

$$\bar{X} = \frac{\sum_{i=1}^N x_i}{N}$$

N es la cantidad de valores a promediar

- Ejemplo

30 36 47 50 52 52 56 60 63 70 70 110

$$\bar{X} = \frac{30 + 36 + 47 + 50 + 52 + 52 + 60 + 63 + 70 + 70 + 110}{12} = \frac{696}{12} = 58$$

> Mediana

- Divide a los valores del atributo en dos partes iguales de manera que los anteriores son todos menores que él y los siguientes son mayores.
- Antes de calcularla deben **ordenarse los valores** del atributo.
- Ejemplo: atributo numérico con una **cantidad impar** de valores

30 36 47 50 52 52 56 57 60 63 70 70 110



$$\tilde{X} = x_{(N+1)/2} = 56$$

> Mediana

- Divide a los valores del atributo en dos partes iguales de manera que los anteriores son todos menores que él y los siguientes son mayores.
- Antes de calcularla deben **ordenarse los valores** del atributo.
- Ejemplo: atributo numérico con una **cantidad par** de valores

30 36 47 50 52 52 56 60 63 70 70 110



$$\tilde{X} = \frac{x_{N/2} + x_{(N+1)/2}}{2} = \frac{52 + 56}{2} = 54$$

> Mediana

- También puede calcularse sobre **atributos ordinales**. En tal caso, el resultado será o bien el valor que divide al conjunto en dos partes iguales o bien se dirá que “la mediana está entre los valores ...”.
- Antes de calcularla deben **ordenarse los valores** del atributo.
- Ejemplo: atributo ordinal con una **cantidad par** de valores

chico	chico	chico	chico	medio	grande	grande	grande
-------	-------	-------	-------	-------	--------	--------	--------



\tilde{X} está entre “chico” y “medio”

> Moda

- La moda es el valor que aparece con mayor frecuencia. Por lo tanto, puede determinarse para atributos cualitativos y cuantitativos.
- En general, un conjunto de datos con dos o más modas es **multimodal**. Si cada valor de los datos ocurre sólo una vez, entonces **no hay moda**.
- Ejemplo: atributo numérico

30	36	47	50	52	52	56	60	63	70	70	110
----	----	----	----	----	----	----	----	----	----	----	-----

- Hay 2 modas y sus valores son 52 y 70
- Ejemplo: atributo nominal

español	inglés	chino	inglés	chino	chino
---------	--------	-------	--------	-------	-------

- La moda es “chino” por ser el valor que aparece más veces

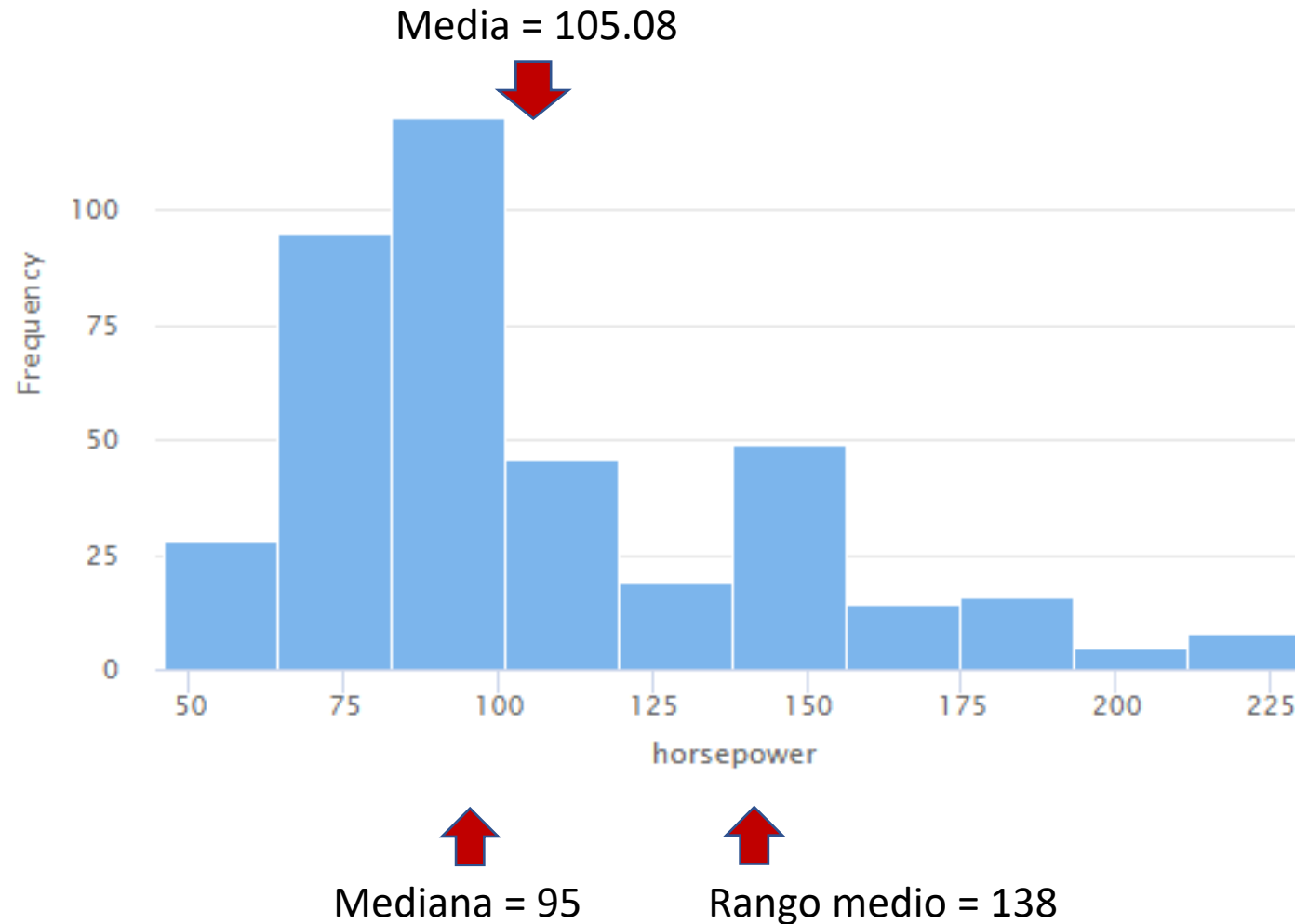
> Rango medio

- El rango medio es fácil de calcular y también puede utilizarse para evaluar la tendencia central de un conjunto de datos numéricos.
- Es la media de los valores máximo y mínimo del conjunto.
- Ejemplo

30 36 47 50 52 52 56 60 63 70 70 110

$$\text{rango medio} = \frac{\text{maximo} + \text{minimo}}{2} = \frac{110 + 30}{2} = \frac{140}{2} = 70$$

Atributo HORSEPOWER

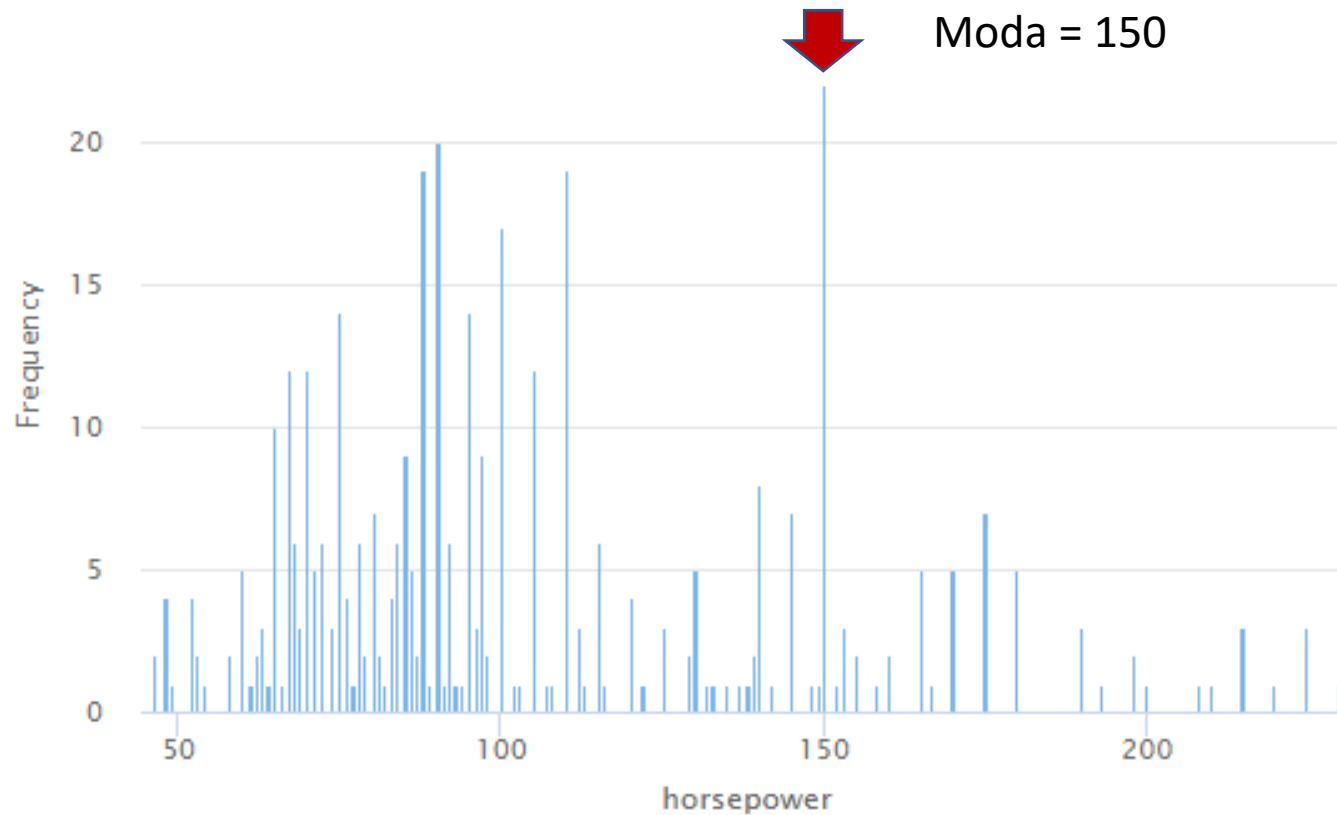


Media	105.08
Mediana	95
Moda	150
Rango medio	138

ID	horsepower
1	46
2	46
...	
199	94
200	95
201	95
202	95
...	
400	225
401	230

Mediana = 95

Atributo HORSEPOWER



Mediana = 95

Media	105.08
Mediana	95
Moda	150
Rango medio	138

ID	horsepower
1	46
2	46
...	
199	94
200	95
201	95
202	95
...	
400	225
401	230

Mediana = 95

Identifican propiedades de los datos y destacan qué valores deben tratarse como ruido o valores atípicos

MEDIDAS DE TENDENCIA CENTRAL

- Media
- Mediana
- Moda
- Rango medio

MEDIDAS DE DISPERSION

- Varianza
- Desviación estándar
- Rango
- Cuartiles
- Rango Inter cuartil

> Varianza y desviación estandar

- La **varianza** mide la dispersión de los datos con respecto a la media. La **desviación estándar** es la raíz cuadrada de la varianza.
- Valores bajos indican que las observaciones de los datos tienden a estar muy cerca de la media, mientras que valores altos indican que los datos están muy dispersos.
- **Estimadores de la varianza muestral**

$$S^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2$$

$$E(S^2) = \frac{n-1}{n} \sigma^2 \quad \text{es sesgado}$$

$$S^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2$$

$$E(S^2) = \sigma^2 \quad \text{es insesgado}$$

> Varianza y desviación estandar

- Ejemplo

30 36 47 50 52 52 56 60 63 70 70 110

- Varianza:

$$S^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2 = \frac{1}{11} ((30 - 58)^2 + (36 - 58)^2 + \dots + (110 - 58)^2)$$

$$S^2 \approx 413.6364$$

- Desviación estándar:

$$S \approx \sqrt{413.6364} \approx 20.3381$$

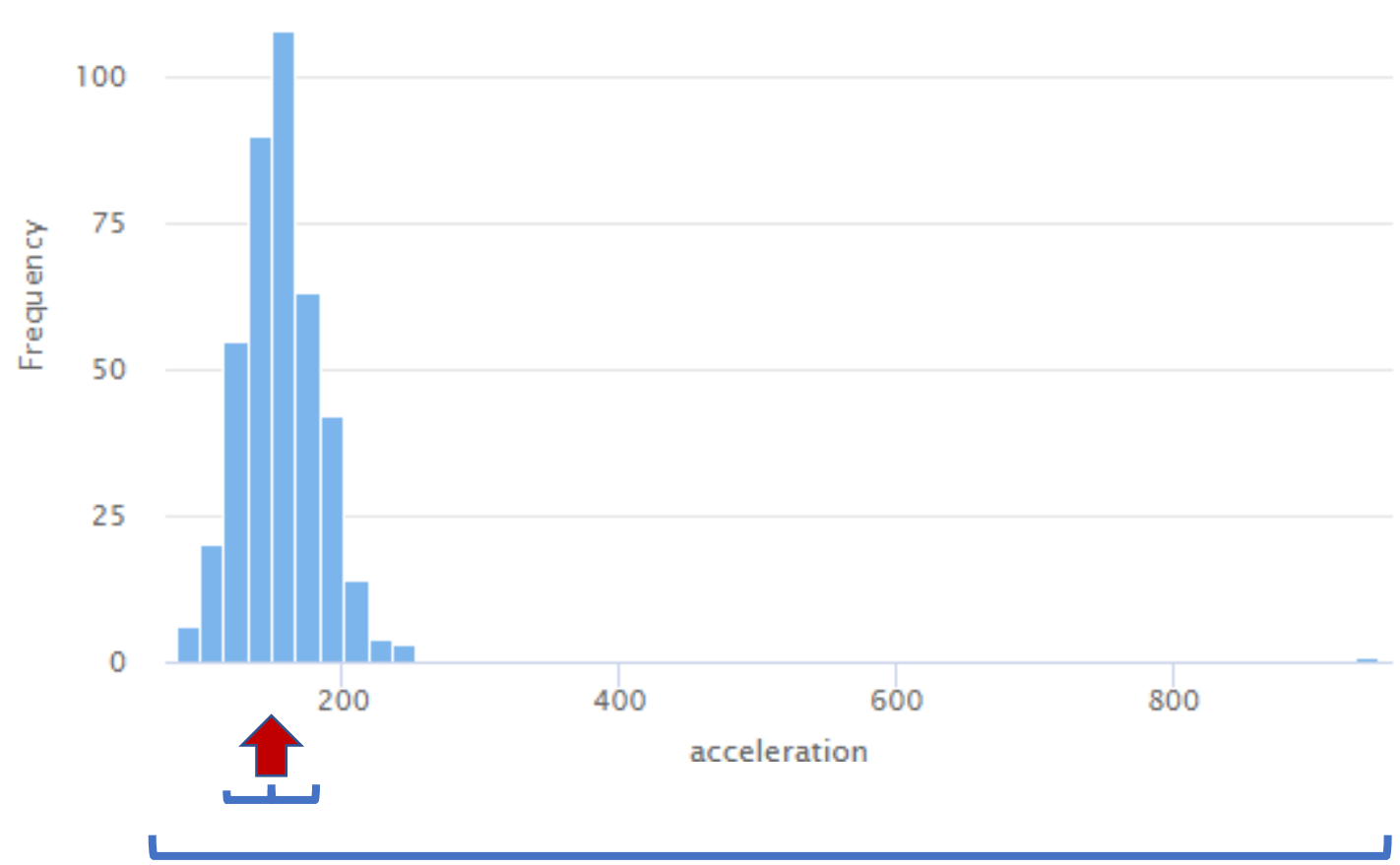
> Rango

- El rango de un conjunto de valores numéricos es la diferencia entre los valores máximo y mínimo de dicho conjunto.
- Ejemplo

30 36 47 50 52 52 56 60 63 70 70 110

$$\text{rango} = \text{maximo} - \text{minimo} = 110 - 30 = 80$$

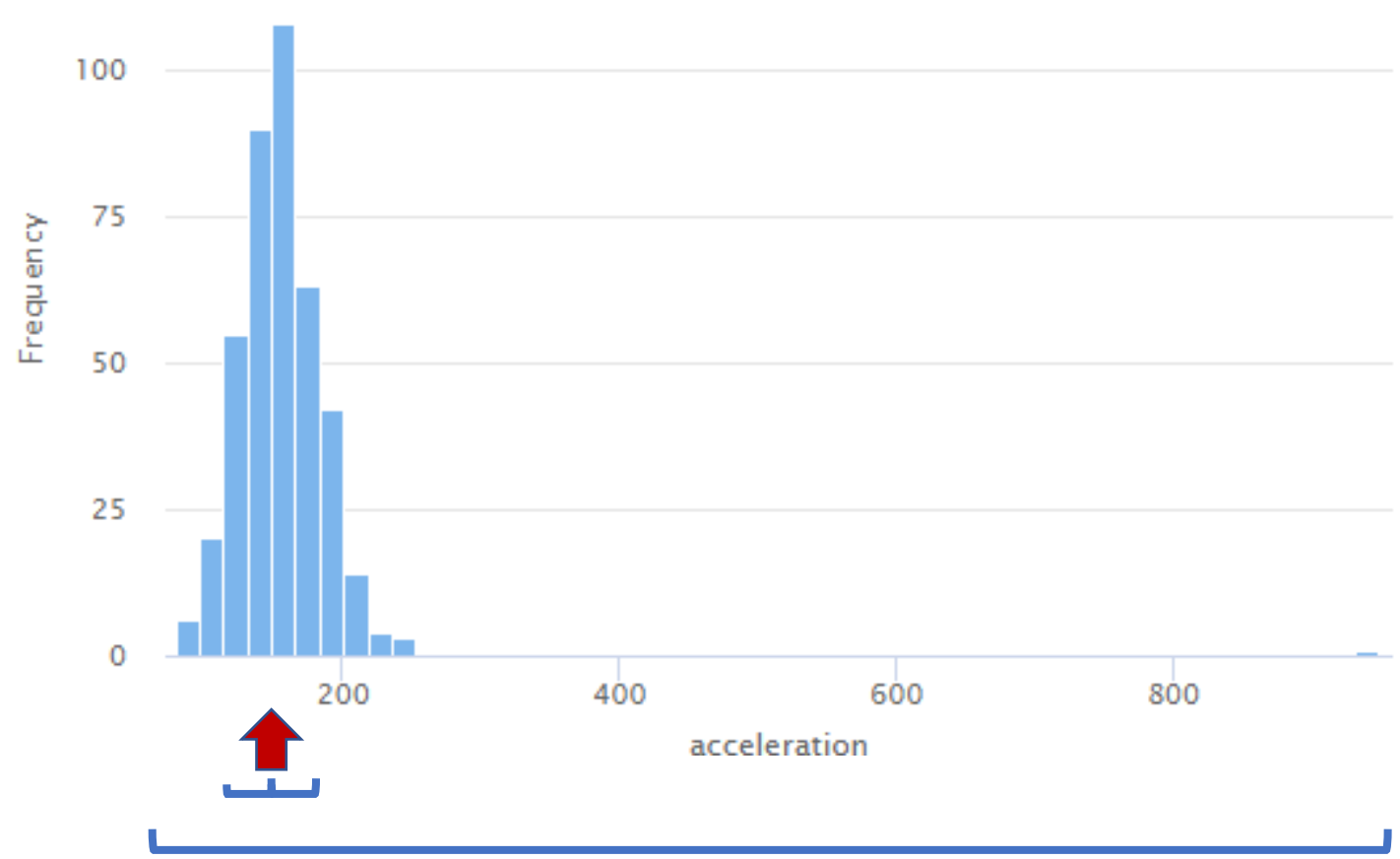
Atributo ACCELERATION



Media	157.30
Desviación	48.29
Minimo	80
Maximo	950
Rango	870

ID	acceleration
...	...
403	237.0
404	246.0
405	248.0
406	950.0

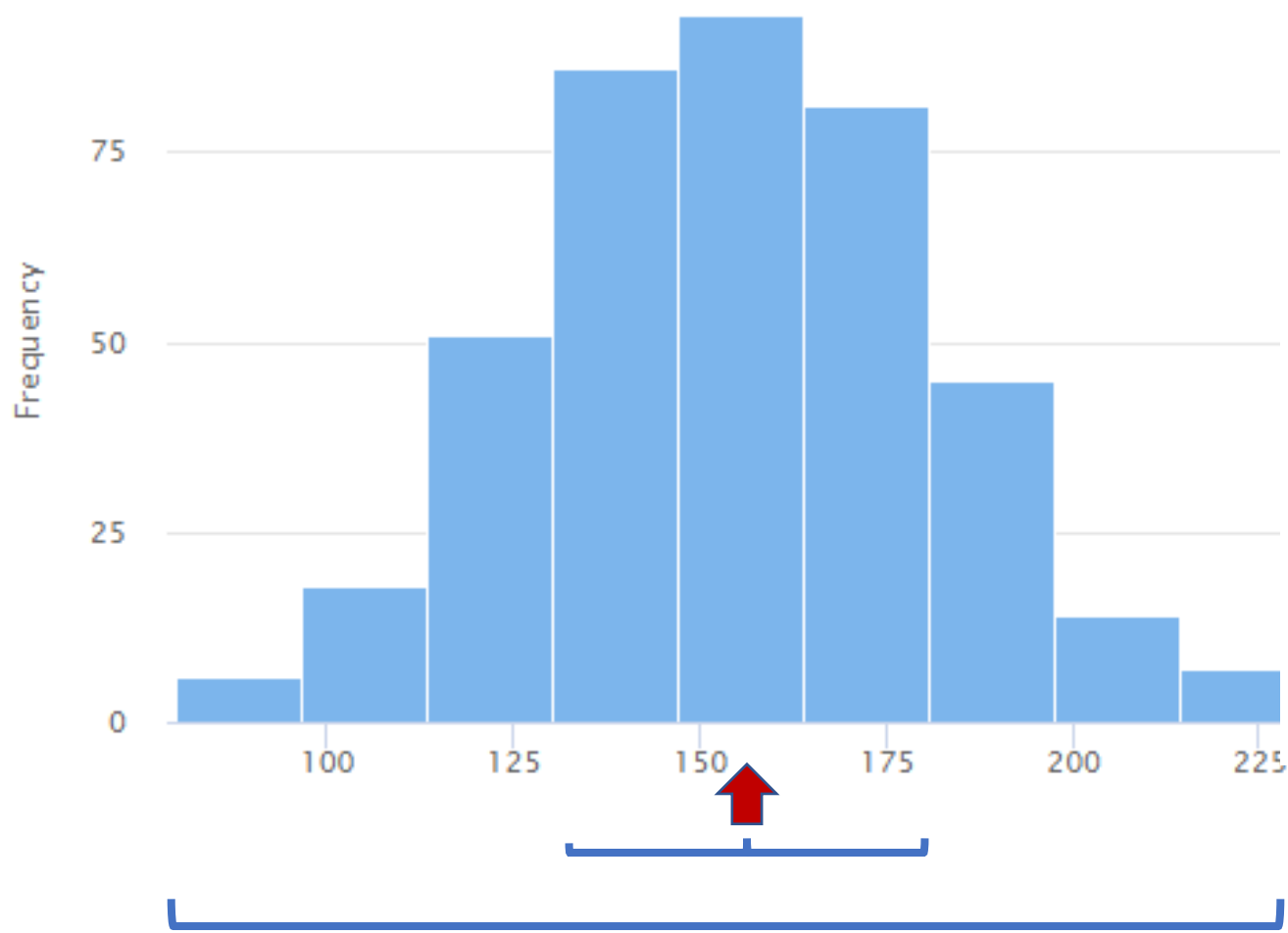
Atributo ACCELERATION



Media	157.30
Desviación	48.29
Minimo	80
Maximo	950
Rango	870

ID	acceleration
...	...
403	237.0
404	246.0
405	248.0
406	950.0

Atributo ACCELERATION



Media	155.35
Desviación	27.91
Minimo	80
Maximo	248
Rango	168

ID	acceleration
...	...
403	237.0
404	246.0
405	248.0
406	950.0

> Cuantiles, cuartiles y percentiles

Los cuantiles son valores que dividen un conjunto numérico ordenado en partes iguales. Es decir que determinan intervalos que comprenden el mismo número de valores.

- CUARTILES: dividen la distribución en cuatro partes.
- DECILES: dividen la distribución en diez partes.
- Centiles o PERCENTILES: dividen la distribución en cien partes.
 - *El percentil es una medida de posición usada en estadística que indica, una vez ordenados los datos de menor a mayor, el valor de la variable por debajo del cual se encuentra un porcentaje dado de observaciones en un grupo.*

30 36 47 50 52 52 56 60 63 70 70 110



$$Q_1 = 47.75$$



$$Q_2 = 54$$



$$Q_3 = 68.25$$


> Rango intercuartil


- La distancia entre Q_1 y Q_3 es una medida sencilla de dispersión que da el rango cubierto por la mitad de los datos.
- Esta distancia se denomina **rango intercuartil (IQR)** y se define como


$$RIC = Q_3 - Q_1$$

- Ejemplo:

30 36 47 50 52 52 56 60 63 70 70 110

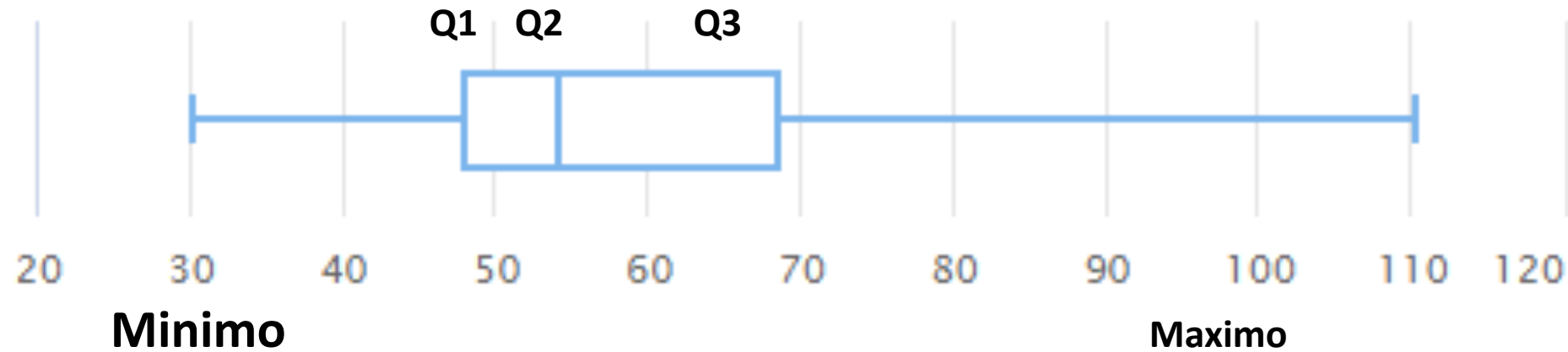

 $Q_1 = 47.75$


 $Q_2 = 54$


 $Q_3 = 68.25$

$$RIC = Q_3 - Q_1 = 68.25 - 47.75 = 20.50$$

> Diagrama de caja simple



El diagrama de caja simple permite analizar la dispersión de los valores de un atributo numérico.

Medidas de dispersion: Dataset Iris

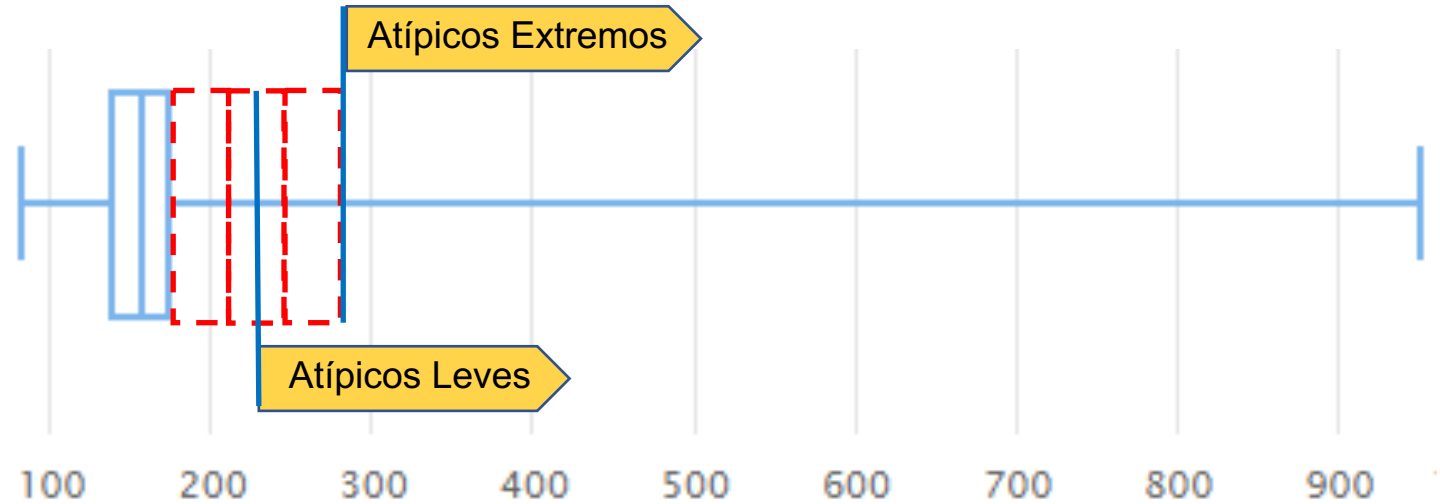


Id	sepalength	sepalwidth	petallength	petalwidth	class
1	5,1	3,5	1,4	0,2	Iris-setosa
2	4,9	3,0	1,4	0,2	Iris-setosa
...
95	5,6	2,7	4,2	1,3	Iris-versicolor
96	5,7	3,0	4,2	1,2	Iris-versicolor
97	5,7	2,9	4,2	1,3	Iris-versicolor
...
149	6,2	3,4	5,4	2,3	Iris-virginica
150	5,9	3,0	5,1	1,8	Iris-virginica



Medidas de dispersion: Atributo Acceleration

Minimo	80
Q1	137
Q2	155
Q3	172.25
Maximo	950

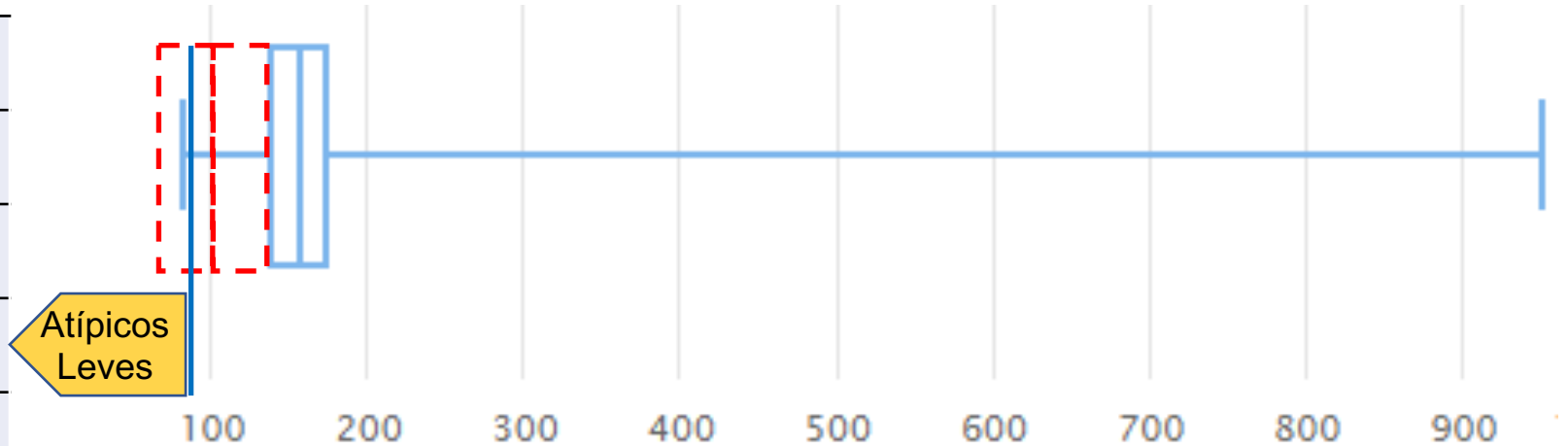


RIC	$Q3 - Q1 = 172.25 - 137 = 35.25$
Lim.Inf	$Q1 - 1.5 * RIC = 137 - 1.5 * 35.25 = 84.125$
Lim.Sup	$Q3 + 1.5 * RIC = 172.25 + 1.5 * 35.25 = 225.125$

¿Hay valores atípicos?

Medidas de dispersion: Atributo Acceleration

Minimo	80
Q1	137
Q2	155
Q3	172.25
Maximo	950



RIC	$Q3 - Q1 = 172.25 - 137 = 35.25$
Lim.Inf	$Q1 - 1.5 * RIC = 137 - 1.5 * 35.25 = 84.125$
Lim.Sup	$Q3 + 1.5 * RIC = 172.25 + 1.5 * 35.25 = 225.125$

¿Hay valores atípicos?

- Los valores de la muestra que pertenezcan a alguno de estos intervalos

$$[Q1 - 3*RIC ; Q1 - 1.5*RIC) \text{ o } (Q3 + 1.5*RIC ; Q3 + 3*RIC]$$

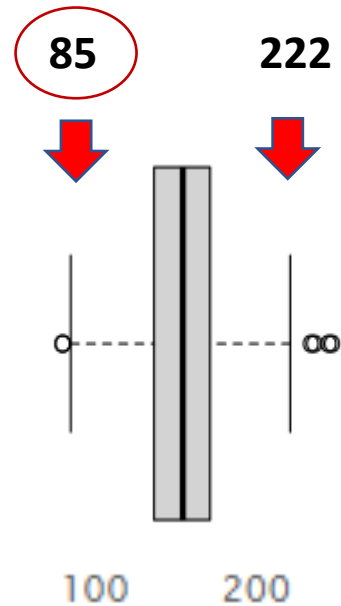
serán considerados **valores atípicos leves**.

- Los valores de la muestra inferiores a

$Q1 - 3*RIC$ o superiores a **$Q3 + 3*RIC$** serán considerados **valores atípicos extremos**.

Diagrama de caja de Tukey

Minimo	80
Q1	137
Q2	155
Q3	172.25
Maximo	950
RIC	35.25
Q1-3*RIC	31.25
Q1-1.5*RIC	84.125
Q3+1.5*RIC	225.125
Q3+3*RIC	278



Los bigotes quedan determinados por los valores del atributo más extremos comprendidos en el intervalo

$$[Q1 - 1.5 * RIC ; Q3 + 1.5 * RIC] = [84.125 ; 225.125]$$

El valor del bigote inferior es el menor valor del atributo que supere $Q1 - 1.5 * RIC$

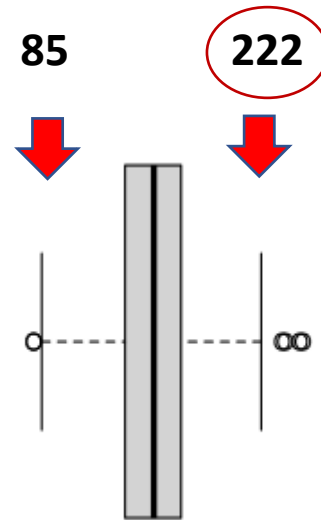
Observando los valores del atributo vemos que el 1er. valor que supera 84.125 es 85

acceleration ↑
80
80
85
85
90
95

Diagrama de caja de Tukey

Minimo	80
Q1	137
Q2	155
Q3	172.25
Maximo	950

RIC	35.25
Q1-3*RIC	31.25
Q1-1.5*RIC	84.125
Q3+1.5*RIC	225.125
Q3+3*RIC	278



100 200 300 400 500 600 700 800 900

Los bigotes quedan determinados por los valores del atributo más extremos comprendidos en el intervalo

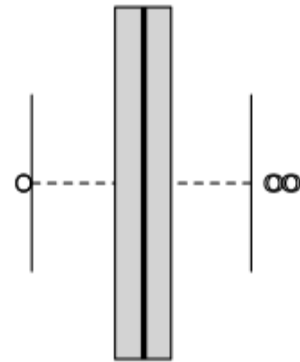
$$[Q1 - 1.5 * RIC ; Q3 + 1.5 * RIC] = [84.125 ; 225.125]$$

El valor del bigote superior es el mayor valor del atributo que no supere $Q3+1.5*RIC$

Observando los valores del atributo vemos que el valor más cercano a 225.125 que no lo supera es 222

acceleration ↑
222
235
237
246
248
950

Minimo	80
Q1	137
Q2	155
Q3	172.25
Maximo	950
<hr/>	
RIC	35.25
Q1-3*RIC	31.25
Q1-1.5*RIC	84.125
Q3+1.5*RIC	225.125
Q3+3*RIC	278



100 200 300 400 500 600 700 800 900

Valor atípico
extremo



- Los valores de ACCELERATION que pertenezcan a **[31.25; 84.125)** o **(225.125; 278]** se considerarán **atípicos leves**.
- Los valores del atributo ACCELERATION inferiores a 31.25 o superiores a 278 se considerarán **atípicos extremos**.

Histograma y diagrama de caja simple

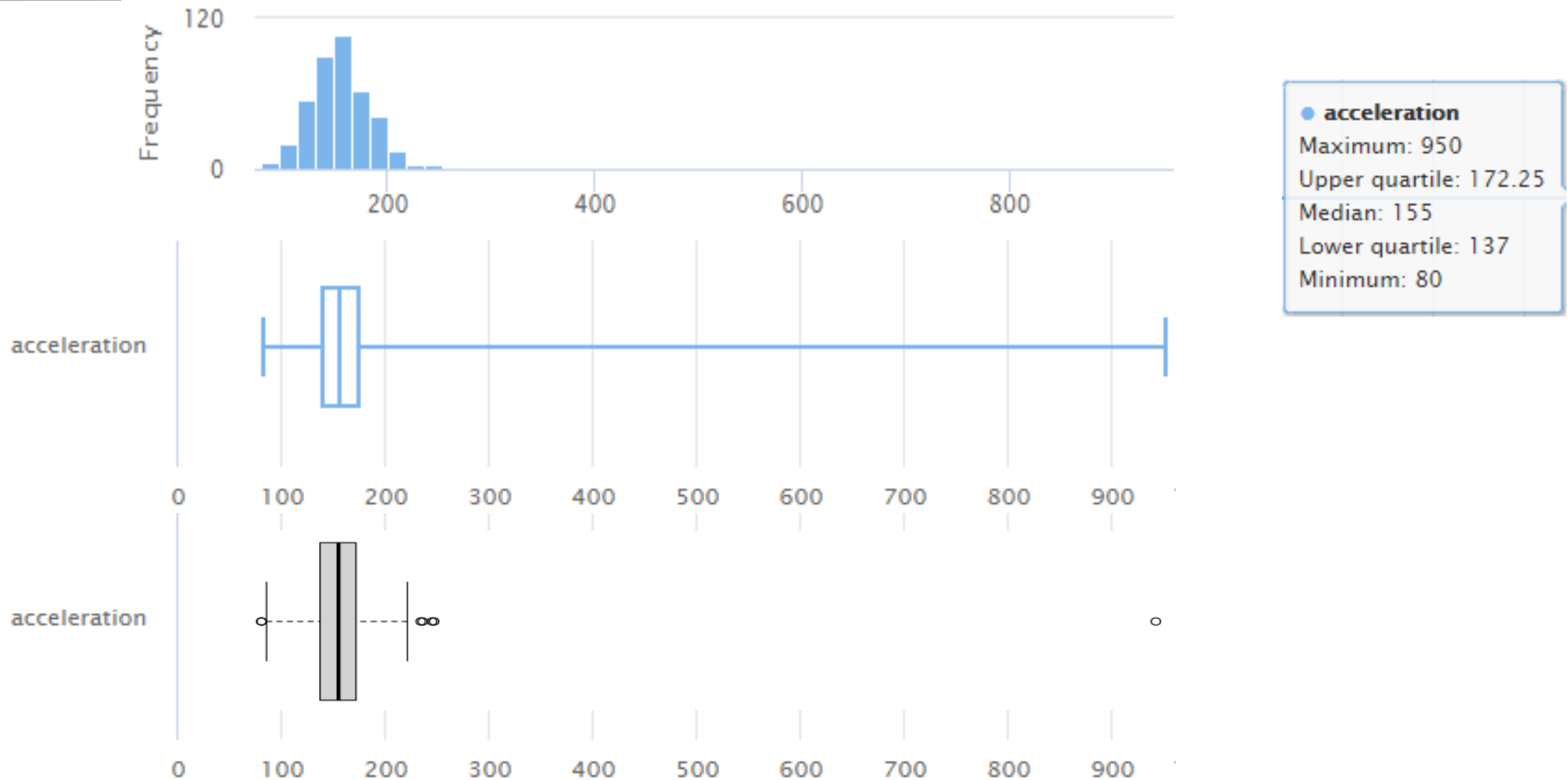
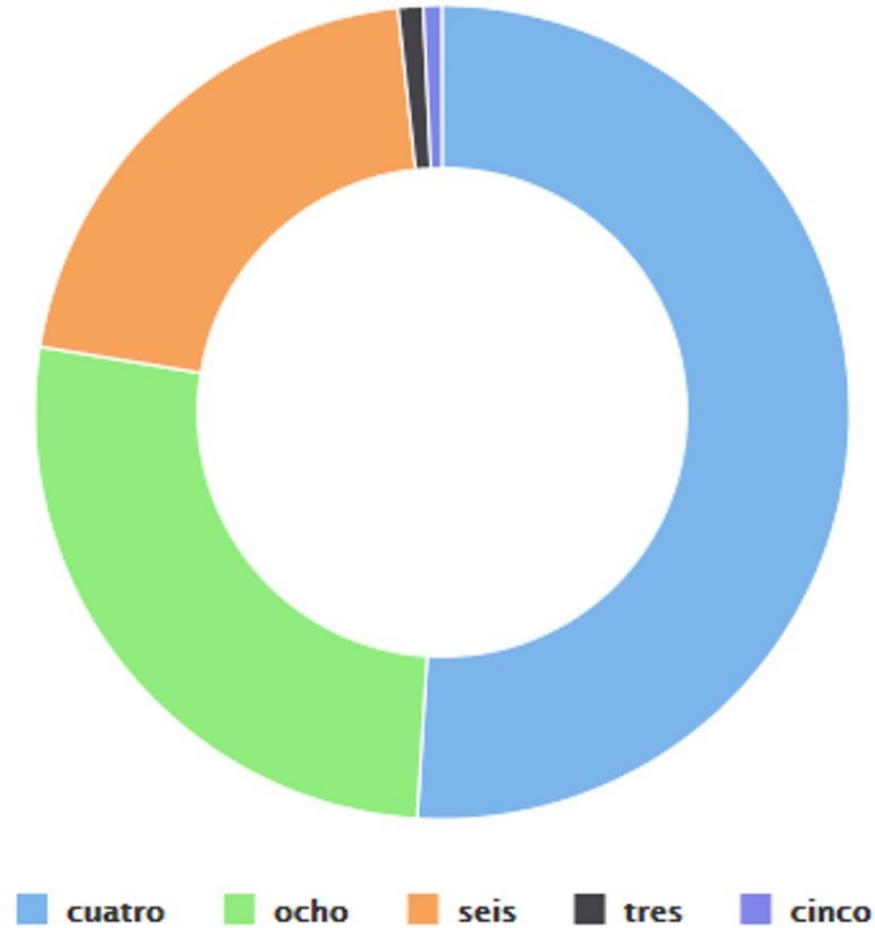
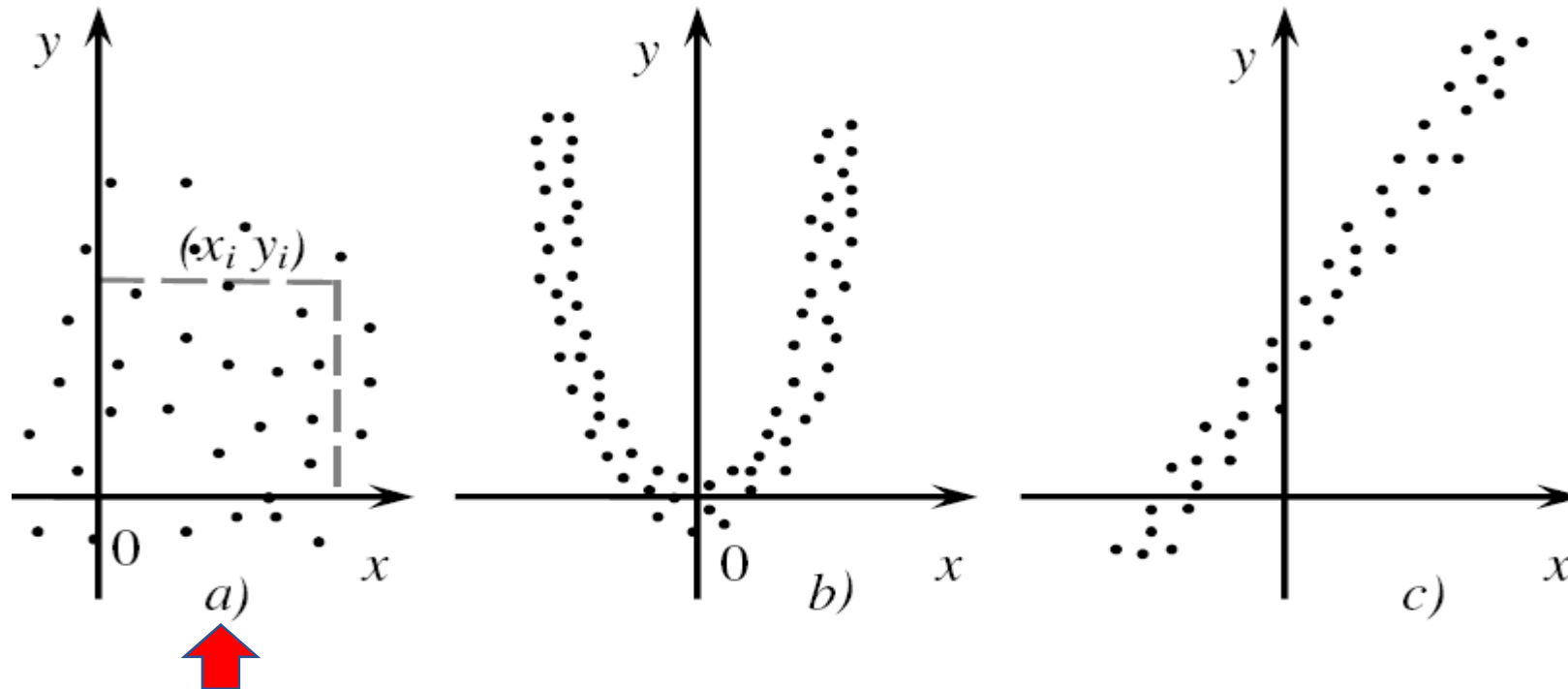


Diagrama de tarta: Attributo Cylinders



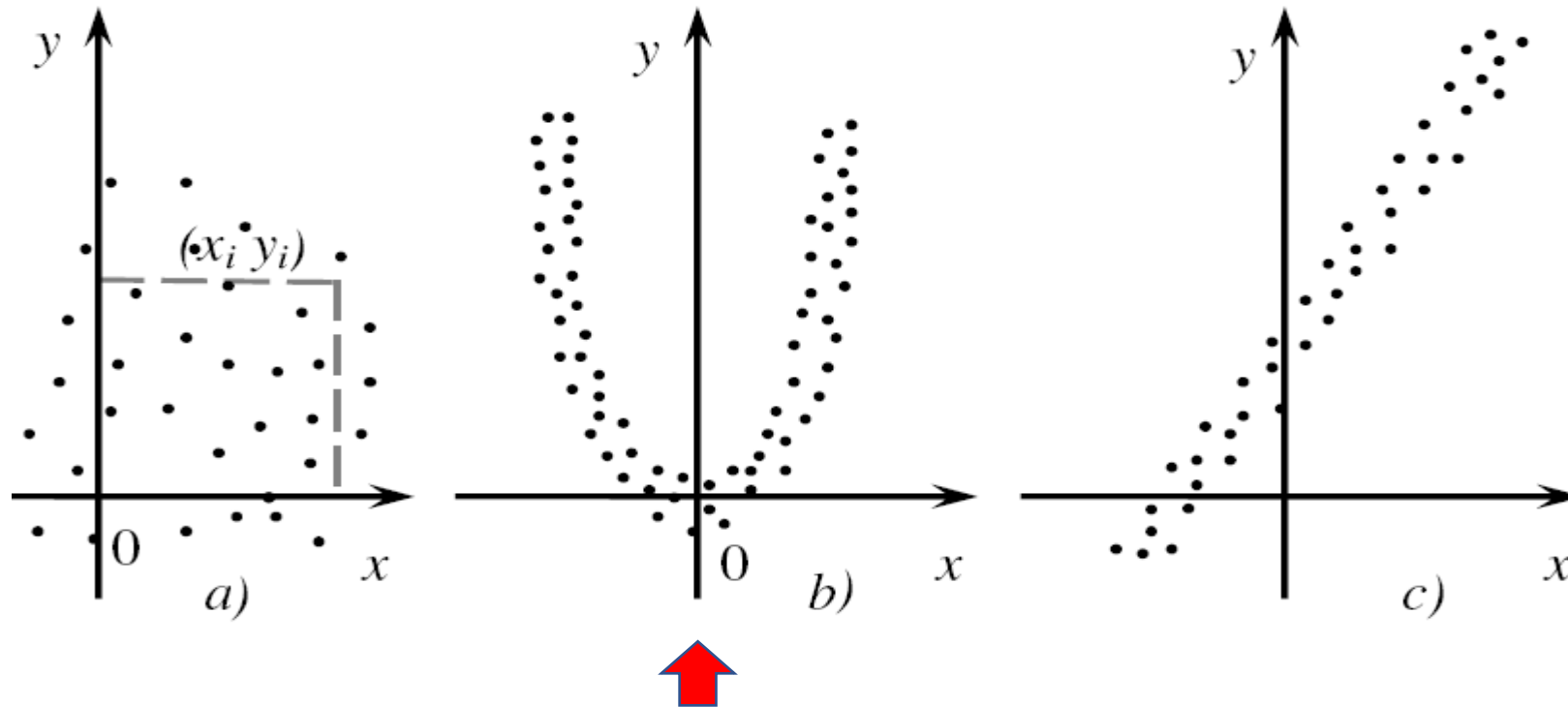
- Consiste en dibujar pares de valores (x_i, y_j) medidos de la v.a. (X,Y) en un sistema de coordenadas



Entre X e Y no hay ninguna relación funcional

Diagrama de dispersión: Atributo horsepower y MPG

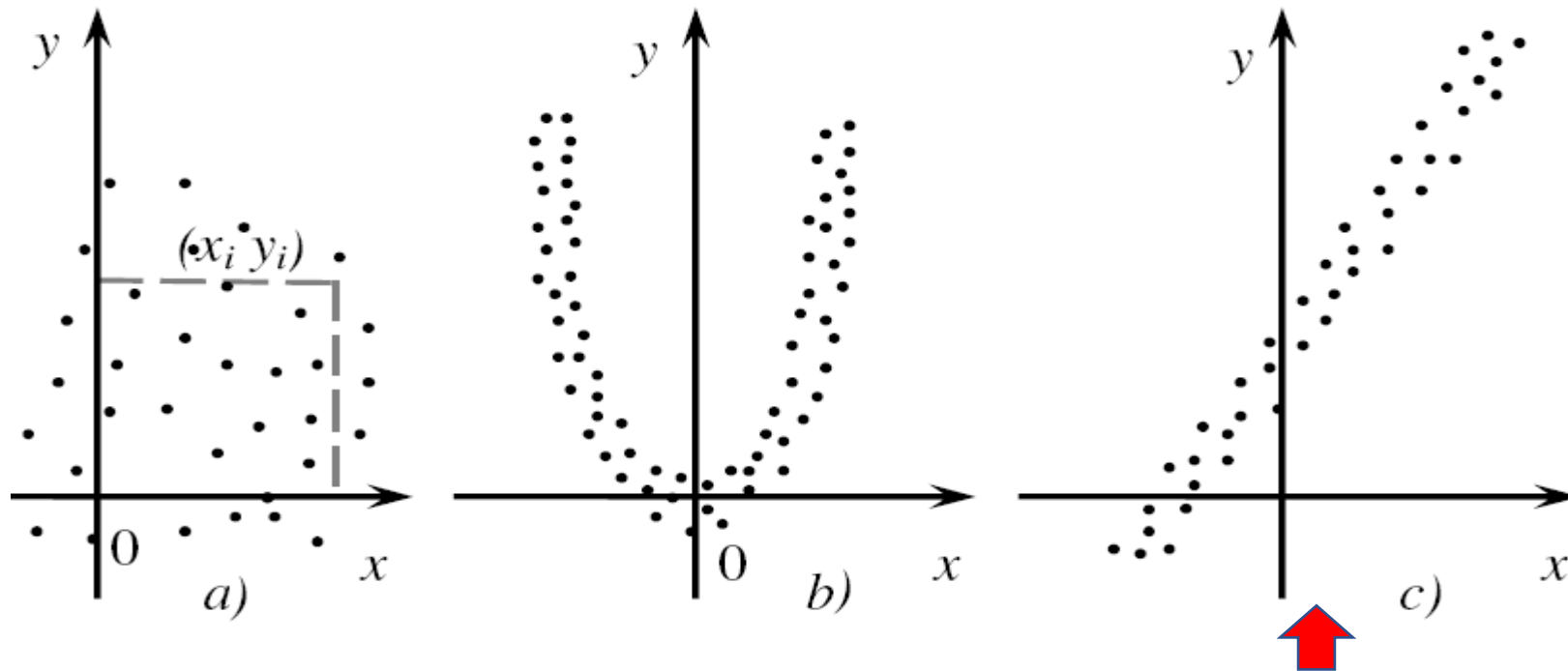
- Consiste en dibujar pares de valores (x_i, y_j) medidos de la v.a. (X,Y) en un sistema de coordenadas



Entre X e Y podría existir una relación funcional que corresponde a una parábola

Diagramas de Dispersión

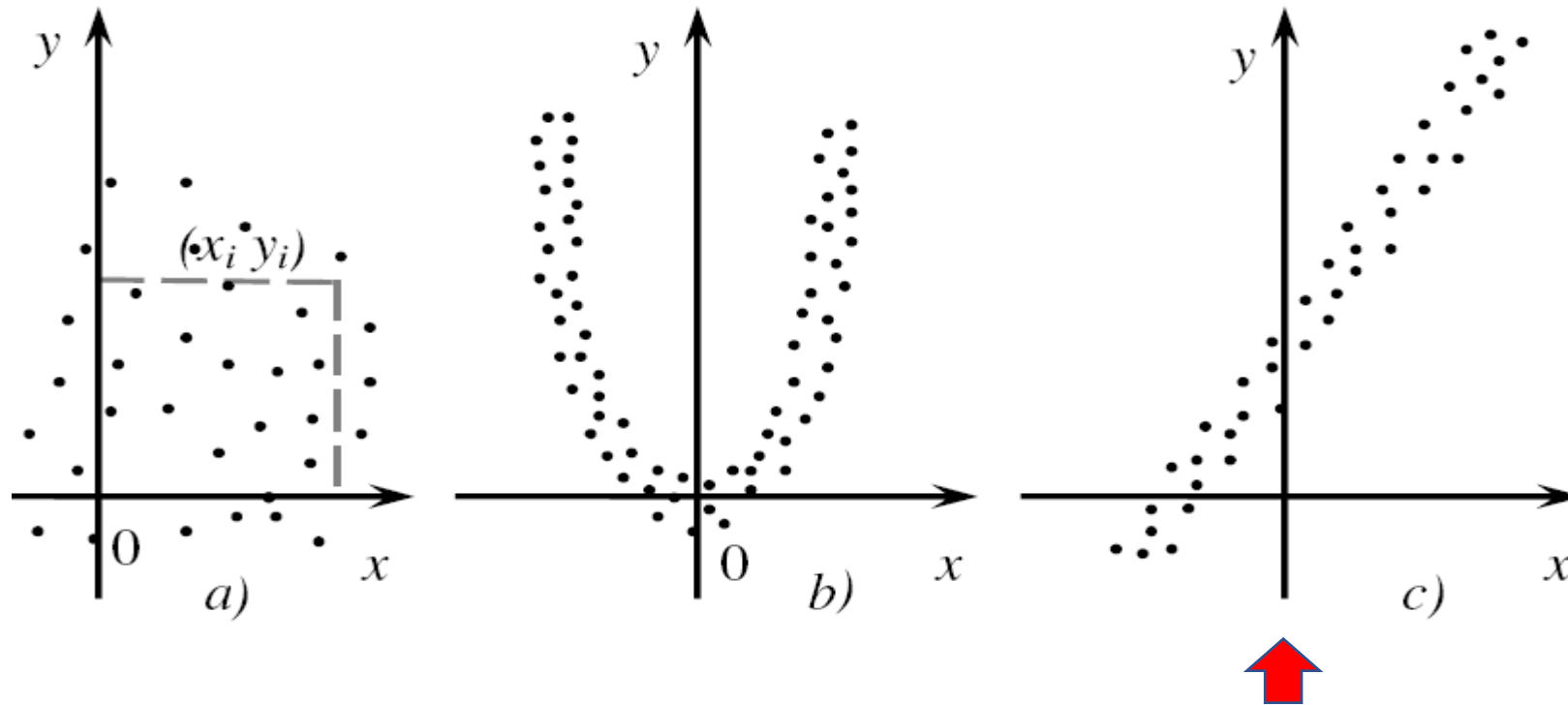
- Consiste en dibujar pares de valores (x_i, y_i) medidos de la v.a. (X,Y) en un sistema de coordenadas



Entre X e Y existe una **relación lineal**. Este es el tipo de relación que nos interesa

Diagrama de dispersión: Atributo horsepower y MPG

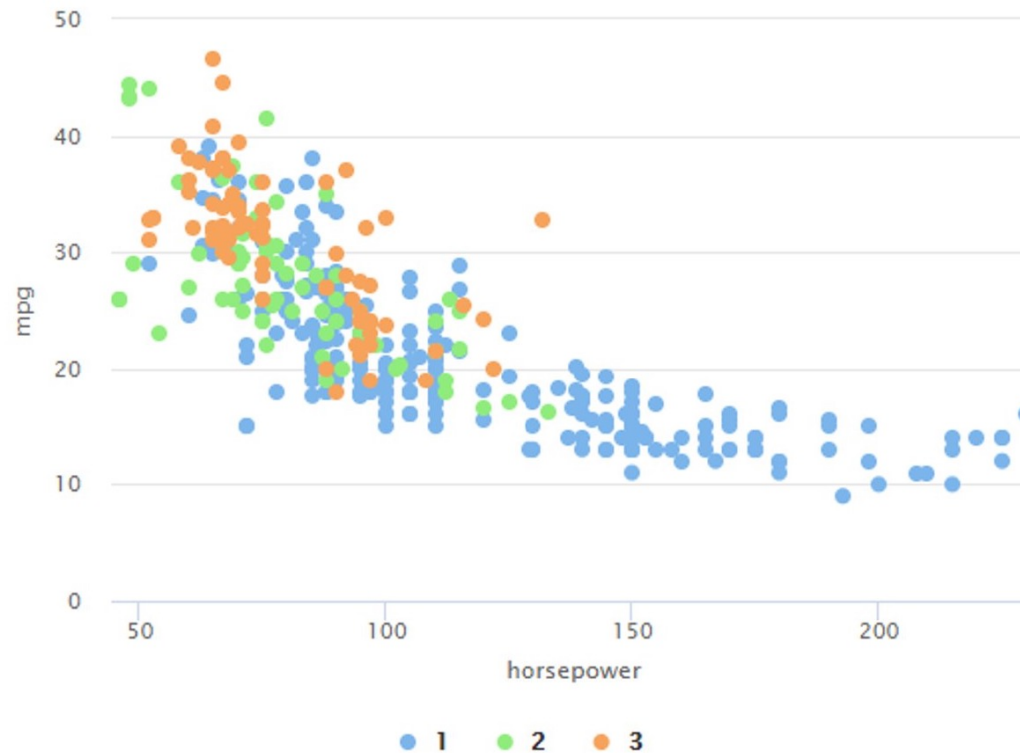
- Consiste en dibujar pares de valores (x_i, y_j) medidos de la v.a. (X,Y) en un sistema de coordenadas



Entre X e Y existe una **relación lineal**. Este es el tipo de relación que nos interesa

Diagrama de dispersión: Atributo horsepower y MPG

- Al momento de construir un modelo de Minería de Datos resulta de interés saber si dos atributos numéricos se encuentran linealmente relacionados o no. Para ello se usa el **coeficiente de correlación lineal**.



INTERPRETACION

- Si $0.5 \leq \text{abs}(\text{Corr}(A,B)) < 0.8$ se dice que A y B tienen una correlación lineal débil.
- Si $\text{abs}(\text{Corr}(A,B)) > 0.8$ se dice que A y B tienen una correlación lineal fuerte
- Si $\text{abs}(\text{Corr}(A,B)) < 0.5$ se dice que A y B no están correlacionados linealmente. Esto NO implica que son independientes, sólo que entre ambos no hay una correlación lineal.

Attributes	mpg	displacement	horsepower	weight	acceleration	model_year
mpg	1	0.402	-0.778	-0.832	0.197	0.579
displacement	0.402	1	-0.291	-0.391	0.107	-0.009
horsepower	-0.778	-0.291	1	0.867	-0.260	-0.424
weight	-0.832	-0.391	0.867	1	-0.183	-0.315
acceleration	0.197	0.107	-0.260	-0.183	1	0.141
model_year	0.579	-0.009	-0.424	-0.315	0.141	1

Para obtener esta matriz todos los atributos deben ser numéricos

Gracias