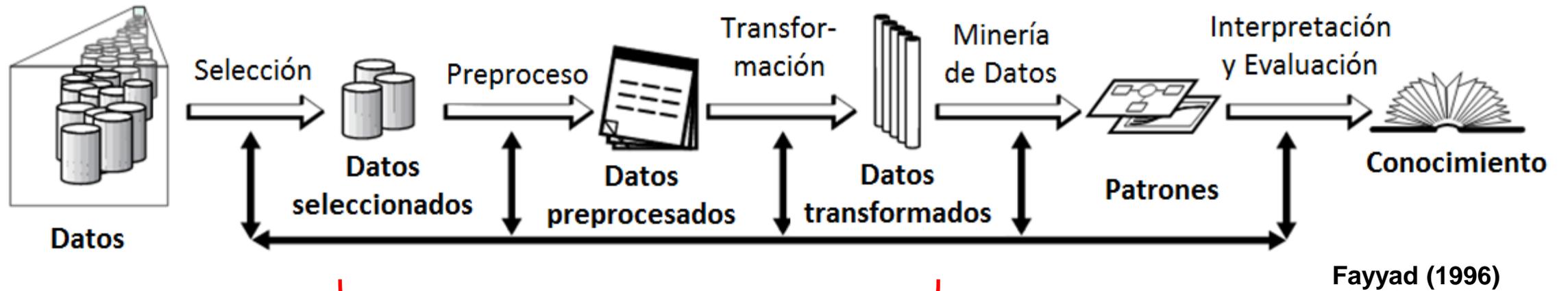


Minería de Datos y el proceso de KDD



Vamos a trabajar en la preparación de los datos para obtener la “*vista minable*”

Fase de Preparación de los Datos

- La información almacenada siempre tiene
 - ▣ Datos faltantes
 - ▣ Valores extremos
 - ▣ Inconsistencias
 - ▣ Ruido
- Tareas a realizar
 - ▣ Limpieza (ej: resolver outliers e inconsistencias)
 - ▣ Transformación (ej: discretización)

Limpieza de los datos

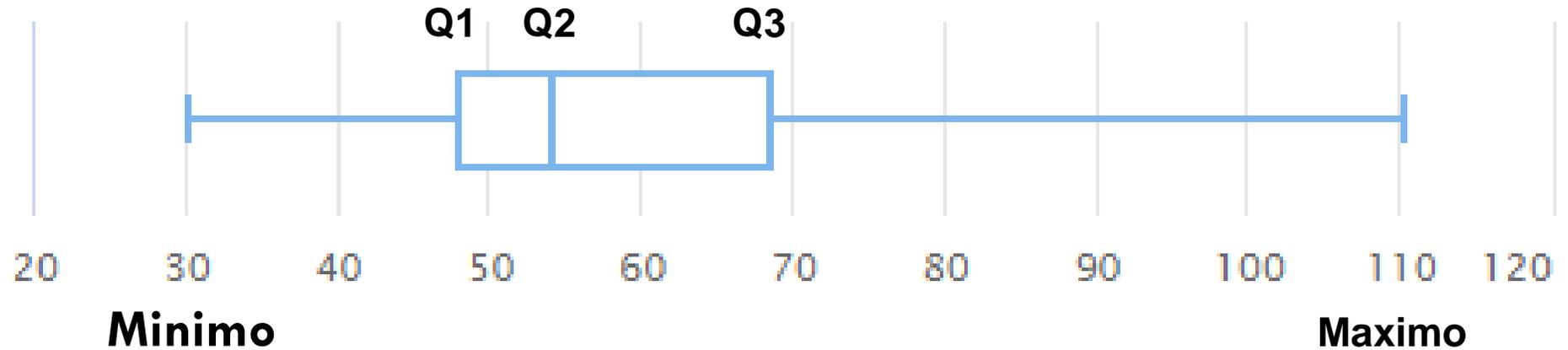
- ✓ En primer lugar, debe tenerse en cuenta que hay distintos tipos de variables o atributos.
- ✓ Para cada tipo se deberá realizar un análisis de sus valores.
 - Luego, se procederá a limpiarlos
 - ▣ Detectar los valores atípicos
 - ▣ Definir qué hacer con los valores faltantes.
 - ▣ Eliminar inconsistencias

Limpieza – Valores atípicos

- Las variables con ruido tendrán valores que caen fuera del rango de sus valores esperados llamados **outliers**.
- Por qué se originan?
 - ▣ Error humano en la carga de datos (ej: una persona puede aparecer con una altura de 5 metros).
 - ▣ Determinados cambios operacionales no han sido registrados en el proceso.

Usaremos “Diagramas de caja y bigotes” como herramienta para detectar valores atípicos

Diagrama de caja simple



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

Flores de 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

Flores de Iris



	sepalength	sepalwidth	petallength	petalwidth
Media	5.84	3.05	3.76	1.20
Desvío	0.83	0.43	1.76	0.76
Mínimo	4.3	2	1	0.1
Q1	5.1	2.8	1.6	0.3
Q2	5.8	3	4.35	1.3
Q3	6.4	3.3	5.1	1.8
Máximo	7.9	4.4	6.9	2.5
RIC	1.3	0.5	3.5	1.5
Rango	3.6	2.4	5.9	2.4

Flores de 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



Cuartiles y RIC del atributo ACCELERATION

- Una vez identificados los cuartiles, puede calcularse el rango intercuartil (RIC)

$Q_1 = 137$ $Q_2 = 155$ $Q_3 = 172.25$

80	...	137	137	...	155	155	155	155	...	172	173	...	950
1	...	101	102	...	202	203	204	205	...	305	306		406

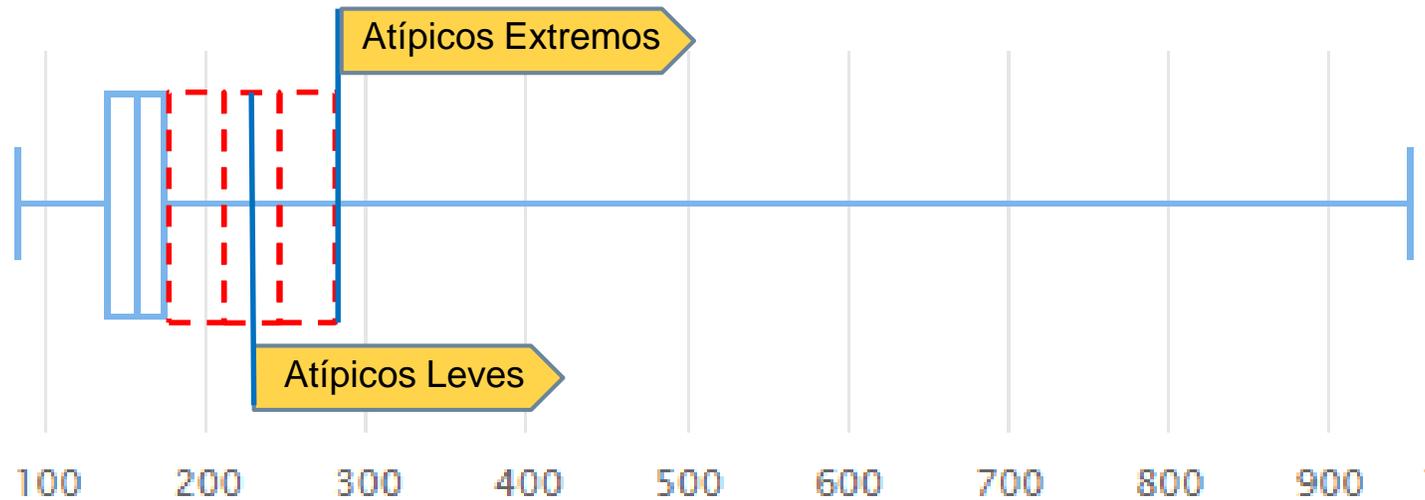
$407/4=101.75$ $3*407/4=305.25$

$$\text{RIC} = Q_3 - Q_1 = 172.25 - 137 = 35.25$$

Diagrama de caja simple

□ Atributo ACCELERATION

Minimo	80
Q1	137
Q2	155
Q3	172.25
Maximo	950



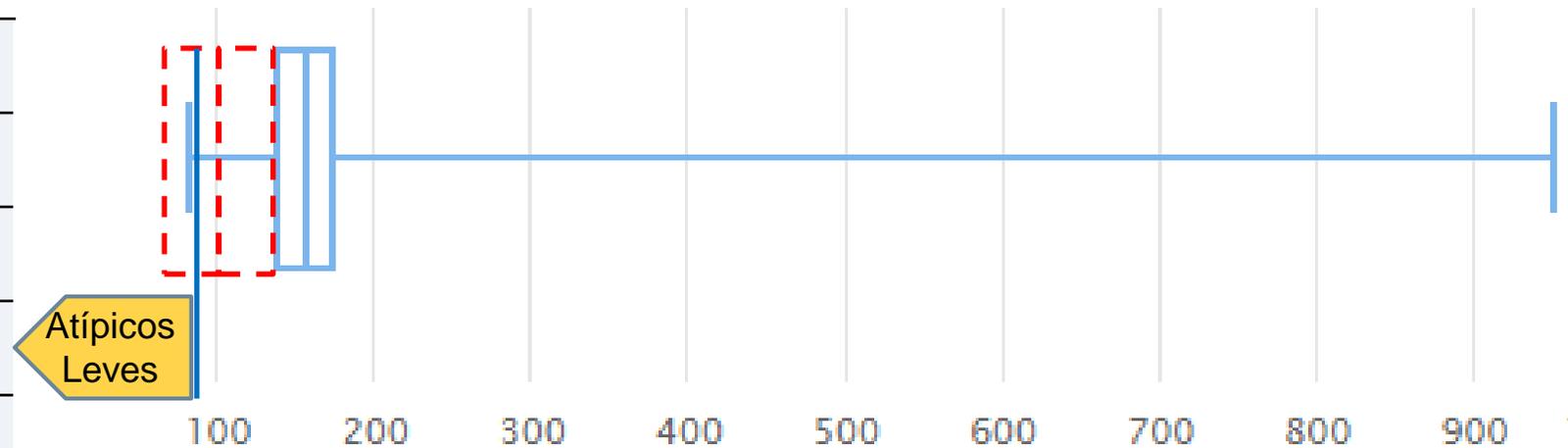
RIC	$Q3 - Q1 = 172.25 - 137 = 35.25$
Lim.Inf	$Q1 - 1.5 * RIC = 137 - 1.5 * 35.25 = 84.125$
Lim.Sup	$Q3 + 1.5 * RIC = 172.25 + 1.5 * 35.25 = 225.125$

¿Hay valores atípicos?

Diagrama de caja simple

□ 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?

Valor atípico o fuera de rango

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

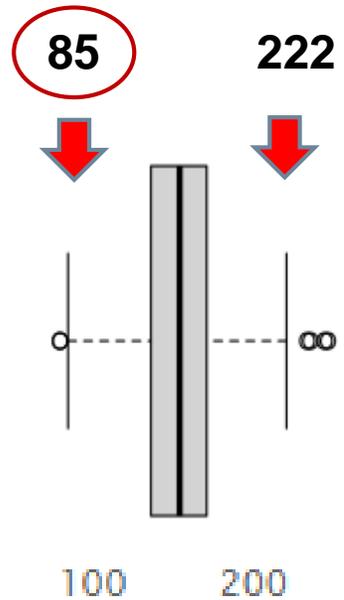
$$[Q1 - 3*RIC ; Q1 - 1.5*RIC) \text{ o } (Q3 + 1.5*RIC ; Q3 + 3*RIC]$$

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



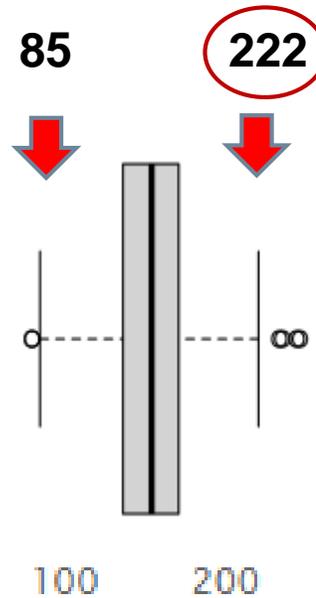
Los bigotes quedan determinados por los valores del atributo más extremos comprendidos en el intervalo
 $[Q1 - 1.5 * RIC ; Q3 + 1.5 * RIC] = [84.125 ; 225.125]$

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



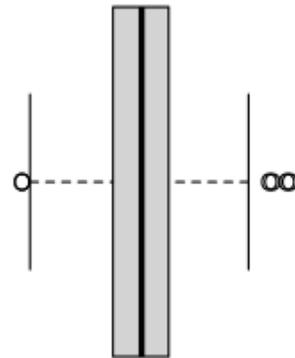
Los bigotes quedan determinados por los valores del atributo más extremos comprendidos en el intervalo
 $[Q1 - 1.5 * RIC ; Q3 + 1.5 * RIC] = [84.125 ; 225.125]$

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

Diagrama de caja de Tukey

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



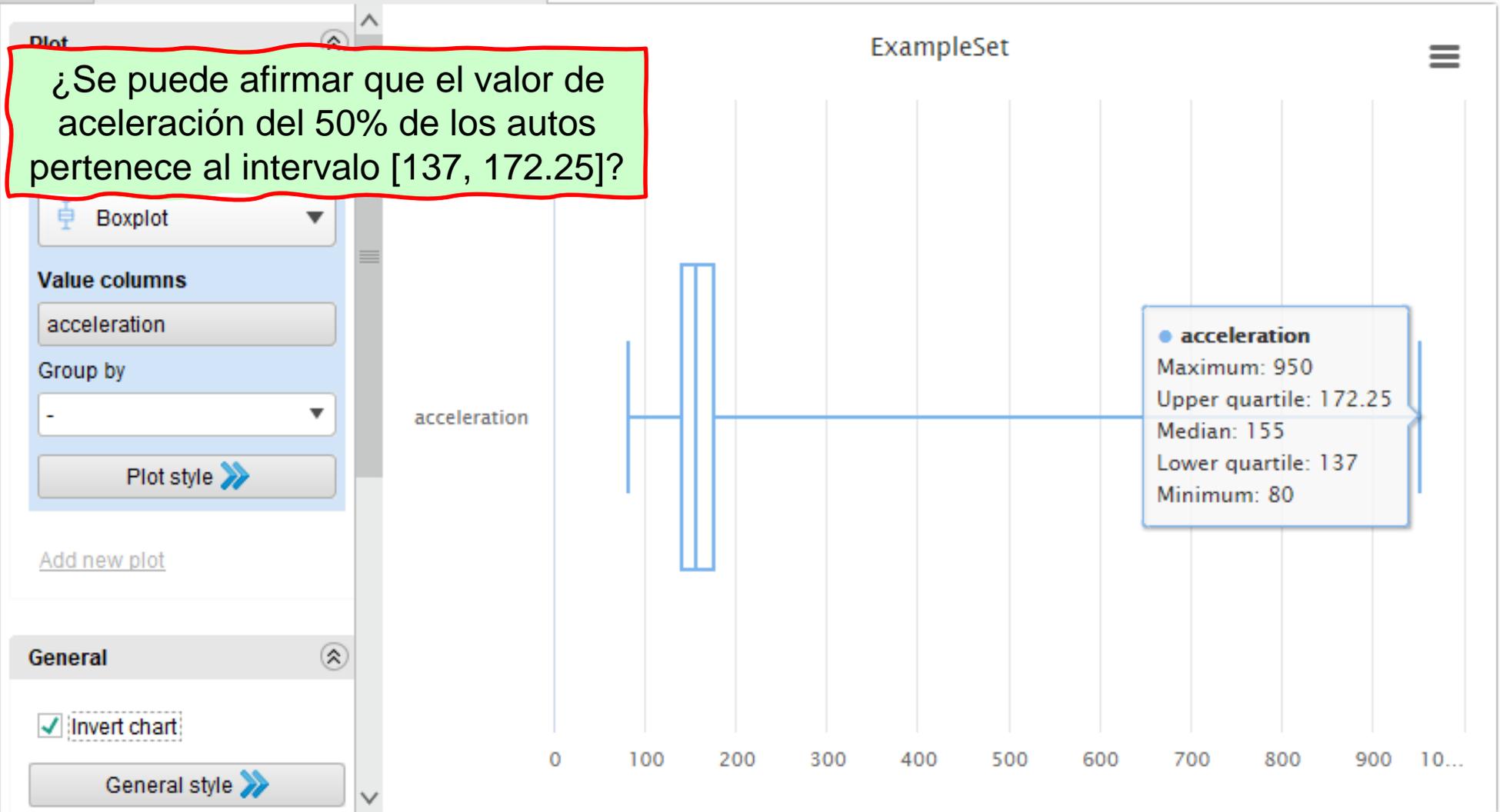
100 200 300 400 500 600 700 800 900

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**.

¿Se puede afirmar que el valor de aceleración del 50% de los autos pertenece al intervalo [137, 172.25]?

- Data
- Statistics
- Visualizations
- Annotations



- Repository
- Imp...
 - Training Resources
 - Samples
 - Community Samples
 - AAP22 (Local)
 - Local Repository
 - MD2022 (Local)
 - Repre_2023 (Local)
 - DB (Legacy)

Data

Statistics

Visualizations

Annotations

Plot

Plot 1

Plot type

Boxplot

Value columns

acceleration

Group by

-

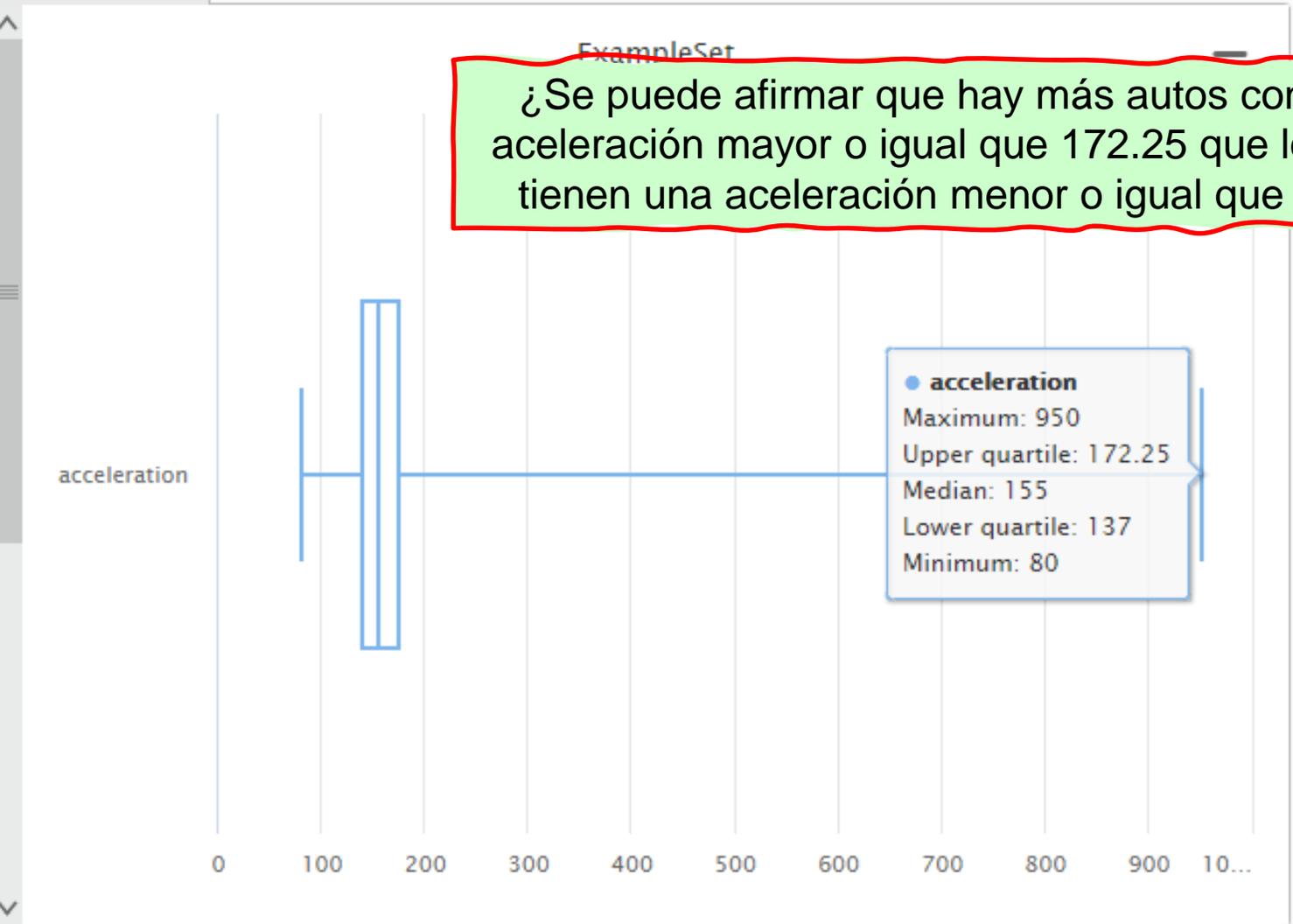
Plot style

Add new plot

General

Invert chart

General style



¿Se puede afirmar que hay más autos con una aceleración mayor o igual que 172.25 que los que tienen una aceleración menor o igual que 137?

Repository

Imp...

AAP22 (Local)

Local Repository

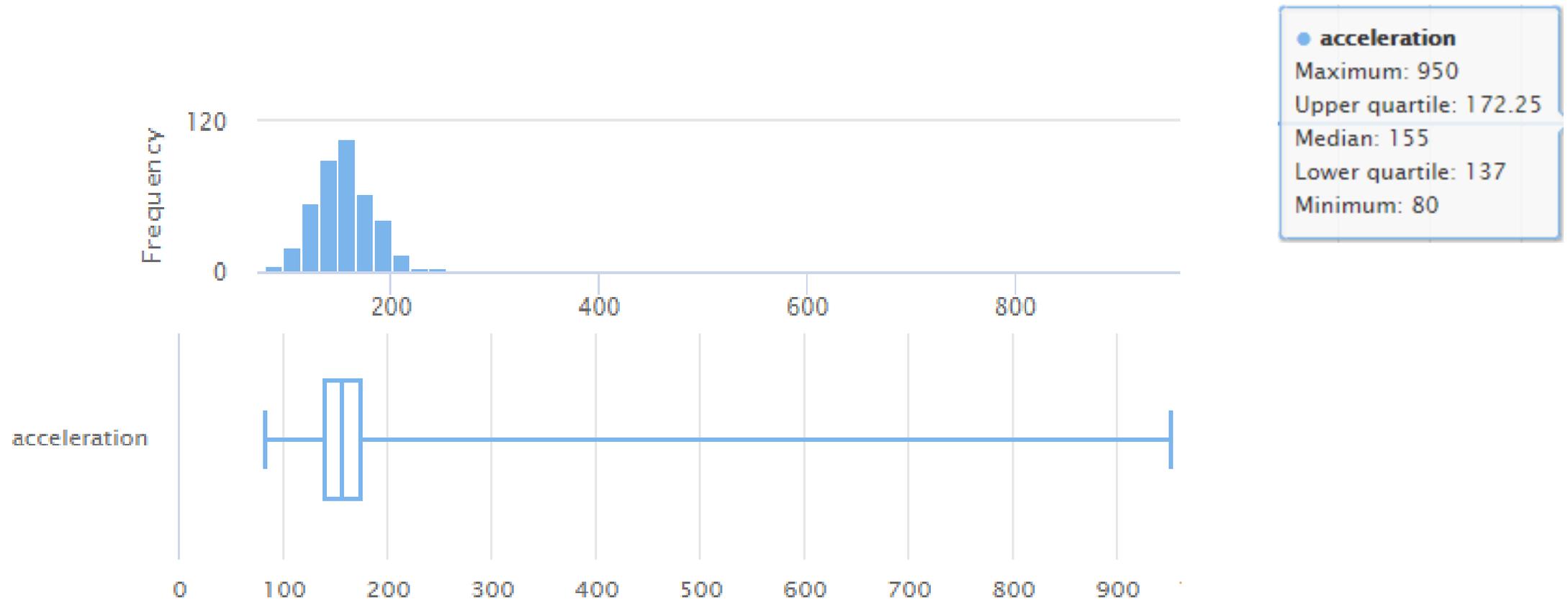
MD2022 (Local)

Repre_2023 (Loc...

DB (Legacy)

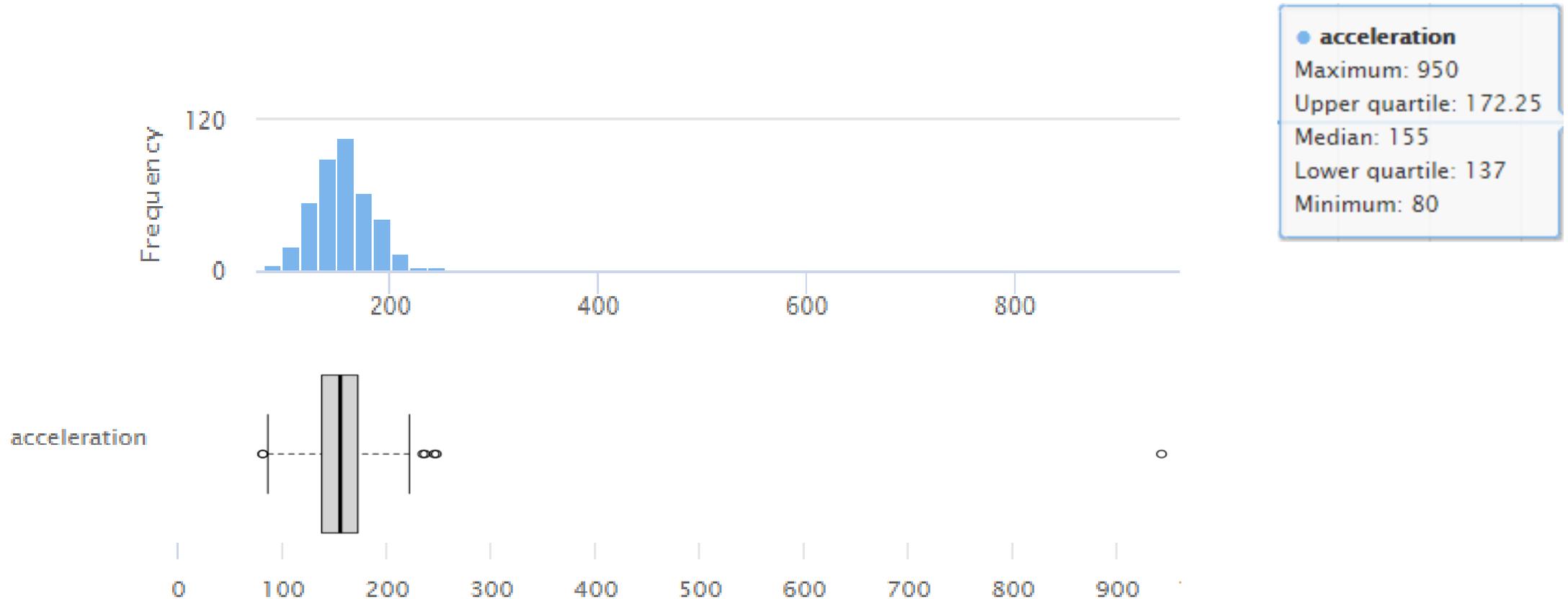
Histograma y diagrama de caja simple

(Atributo ACCELERATION archivo autos-mpg.csv)



Histograma y diagrama de caja de Tukey

(Atributo ACCELERATION archivo autos-mpg.csv)



Limpieza - Valores faltantes

- Qué hacer con los valores nulos?
 - ▣ Ignorar la tupla.
 - ▣ Rellenar la tupla manualmente.
 - ▣ Usar una constante global para rellenar el valor nulo.
 - ▣ Utilizar el valor de la media u otra medida de centralidad para rellenar el valor.
 - ▣ Utilizar el valor de la media u otra medida de centralidad de los objetos que pertenecen la misma clase.
 - ▣ Utilizar alguna herramienta de Minería de Datos para calcular el valor más probable.

Views: Find data, operators...etc All Studio

Hay datos faltantes
¿cómo los completamos?

Result History **ExampleSet (autos-mpg.csv)**

Name	Type	Missing	Filter (9 / 9 attributes):	Search for Attributes
mpg	Real	8	Min 9	Max 46.600
cylinders	Nominal	0	Least cinco (3)	Most cuatro (207)
displacement	Integer	0	Min 1000	Max 9800
horsepower	Real	6	Min 46	Max 230
weight	Integer	0	Min 1613	Max 5140
			Min	Max

Showing attributes 1 - 9 Examples: 406 Special Attributes: 0 Regular Attributes: 9

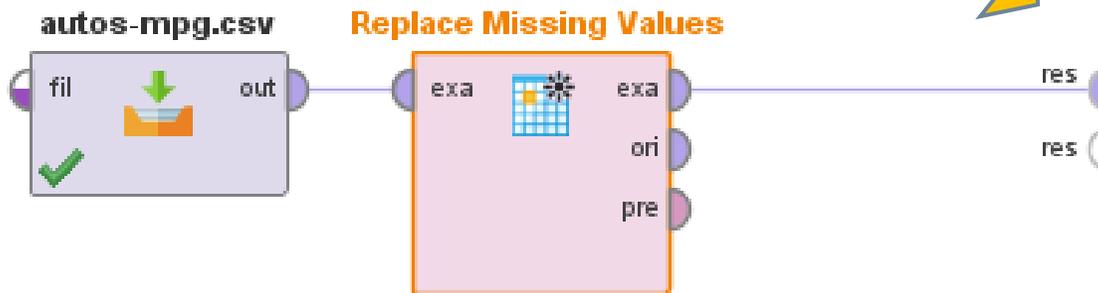
Repository

- Training Resources
- Community Samples
- Samples
- AAP22 (Local)
- Local Repository
- MD2022 (Local)
- Repre_2023 (Local)
- DB (Legacy)

Reemplazando los valores faltantes

The screenshot displays the RapidMiner Studio interface. The title bar reads "<new process*> - RapidMiner Studio Free 9.10.013 @ DESKTOP-BBJS360". The menu bar includes File, Edit, Process, View, Connections, Settings, Extensions, and Help. The toolbar contains icons for file operations and a 'Views' dropdown set to 'Design'. On the left, the 'Operators' panel is open, showing a search for 'missi' and a list of categories: Blending (1), Examples (1), Filter (1), Cleansing (9), and Missing (8). The 'Replace Missing Values' operator is highlighted with a blue selection box, and a yellow arrow points to it from a text box. The main 'Process' canvas shows a workflow starting with an input 'inp' connected to a 'Process' operator. The operator is labeled 'autos-mpg.csv' and has 'fil' and 'out' ports. A green checkmark is visible on the operator icon. The output 'out' is connected to a 'res' output port. A yellow text box at the bottom right contains the text 'Replace Missing Values'.

Todos los atributos se completarán con el promedio o la moda



Parameters

Replace Missing Values

create view

attribute filter type: all

invert selection

include special attributes

default: average

Edit Parameter List: columns

List of replacement functions for each column.

attribute	replace with
cylinders	average
mpg	minimum

Add Entry Remove Entry Apply Cancel

Se pueden indicar reemplazos específicos

Change compatibility (9.10.013)



EJECUTE y verifique cómo se completaron los atributos mpg y horsepower

- Result History
- Example
- Data
- Statistics
- Visualizations
- Simple Charts

Name	Type	Missing	Filter (9 / 9 attributes):	Min	Max
mpg	Real	0	Search for Attributes	9	46.600
cylinders	Polynomial	0		Least cinco (3)	Most cuatro (207)
displacement	Integer	0		1000	9800
horsepower	Real	0		46	230
weight	Integer	0		1613	5140
acceleration	Integer	0		80	950

Showing attributes 1 - 9 Examples: 406 Special Attributes: 0 Regular Attributes: 9

- Repository
- Import...
 - Training Resource
 - Community Samp
 - Samples
 - AAP22 (Local)
 - Local Repository (t
 - MD2022 (Local)
 - Repre_2023 (Loca
 - DB (Legacy)



Transformación de atributos

- Es una de las etapas más importantes porque de ella depende el éxito del proceso.
- Los atributos serán transformados según las necesidades del algoritmo a aplicar.
- Es probable que deban derivarse variables nuevas.
- También es posible que se reduzcan variables convirtiéndolas en información más significativa.

Transformación de atributos

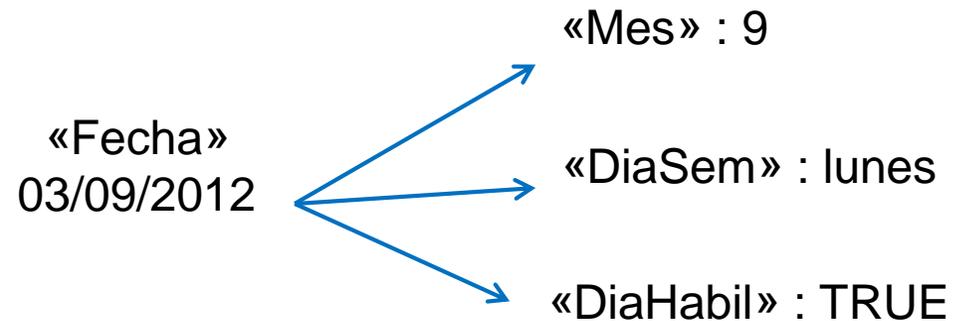
- **Según el algoritmo a aplicar**, las transformaciones más habituales son:
 - ▣ Reducción de dimensionalidad
 - ▣ Aumento de dimensionalidad
 - ▣ Discretización de atributos numéricos
 - ▣ Numerización de atributos nominales
 - ▣ Normalización de atributos

Transformación de atributos

- Reducción de dimensionalidad
 - ▣ Cambia el espacio de entrada por otro que tiene menor dimensión.
 - ▣ Se busca mejorar la relación entre la cantidad de ejemplos y la cantidad de atributos.
 - ▣ **Ejemplos**
 - Análisis de componentes principales (PCA)
 - Red SOM (self-organizing maps)

Transformación de atributos

- Aumento de la dimensionalidad a través de la **creación de características**
 - Atributos numéricos : se utiliza suma, resta, producto, división, máximo, mínimo, media, cuadrado, raíz cuadrada, seno, coseno, etc.
 - Fechas: brindan poca información si se las usa directamente.



Transformación de atributos

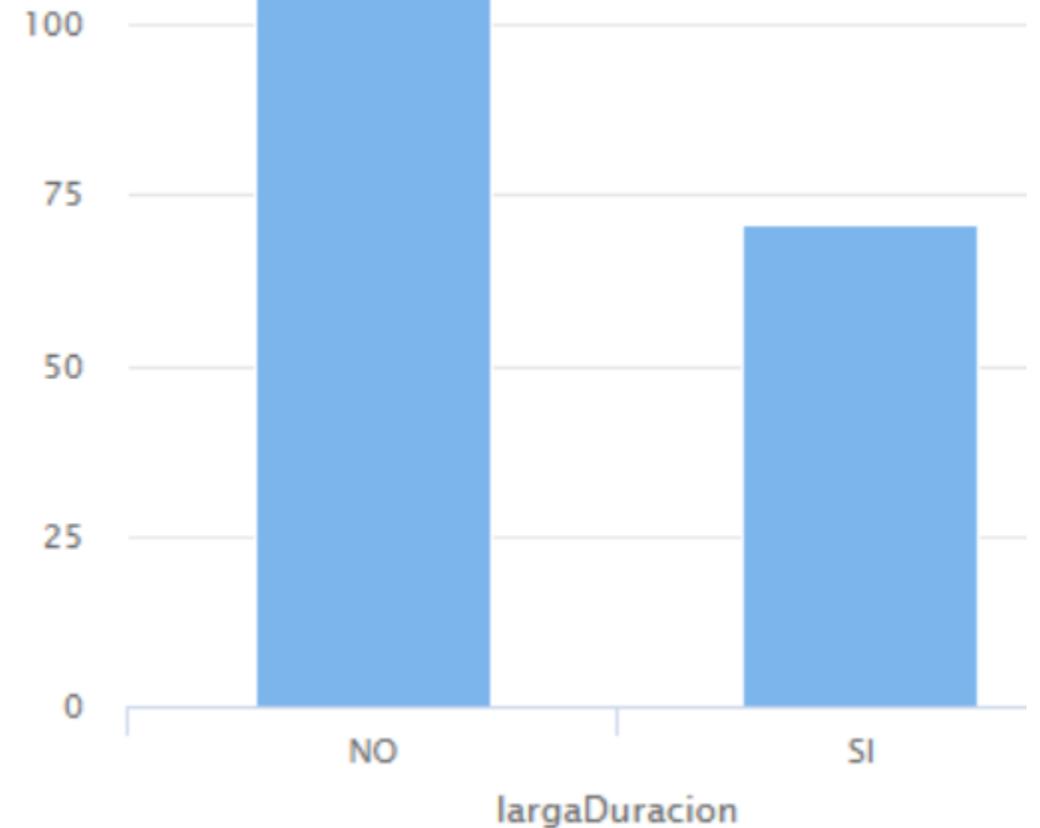
- Aumento de la dimensionalidad a través de la **creación de características**
 - Atributos nominales:
 - Se utilizan las operaciones lógicas, igualdad o desigualdad, condiciones **M-de-N** (TRUE si al menos M de las N condiciones son verdaderas).
 - Se puede generar un valor numérico a partir de valores nominales, por ejemplo, las variables **X-de-N** (retorna el entero X de las N condiciones que son ciertas)

Ejemplo de creación de atributos

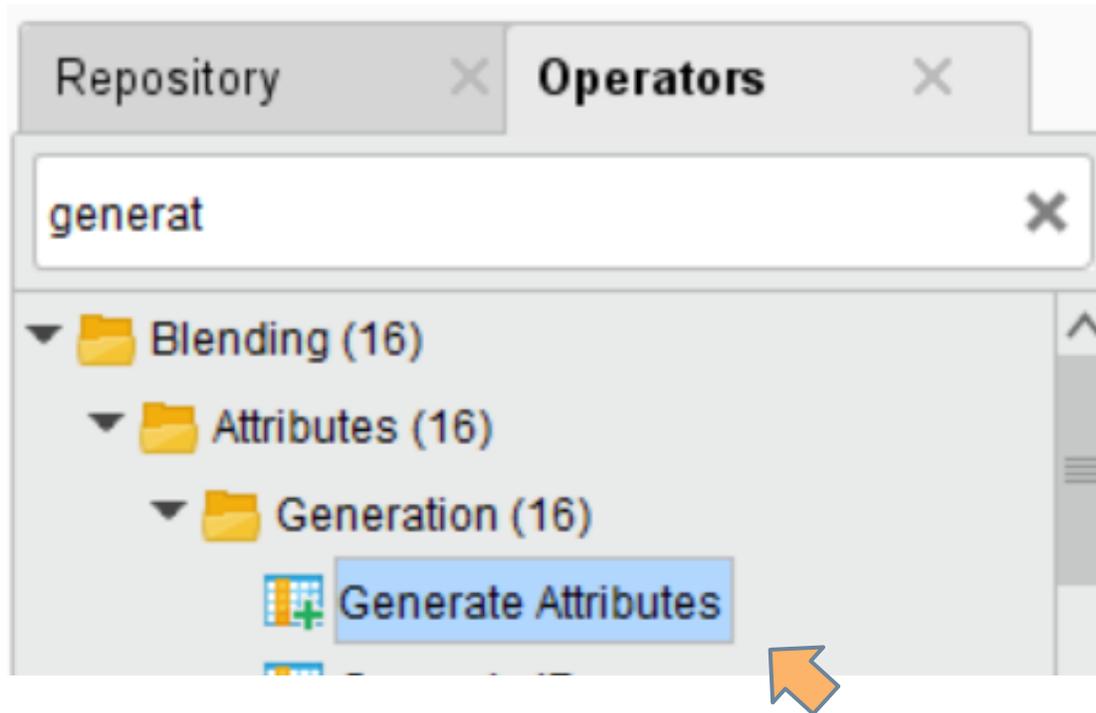
Atributo derivado	Fórmula
Índice de obesidad	$\text{Altura}^2 / \text{peso}$
Hombre familiar	Casado, varón e (hijos > 0)
Síntomas SARS	3-de-5 (fiebre alta, vómitos, tos, diarrea, dolor de cabeza)
Riesgo de póliza	X-de-N (edad < 25, varón, años que conduce < 2, vehículo deportivo)
Beneficios Brutos	Ingresos – Gastos
Beneficios netos	Ingresos – Gastos – Impuestos
Desplazamiento	Pasajeros * kilómetro
Duración media	Segundos de llamada / número de llamadas
Densidad	Población / Area
Retardo compra	Fecha compra – Fecha campaña

Ejercicio

- Genere un nuevo atributo **DECADA** cuyo valor será “70s” si el **modelo** del auto es <80 y “80s” si no.
- Grafique este nuevo atributo utilizando un diagrama de barras.



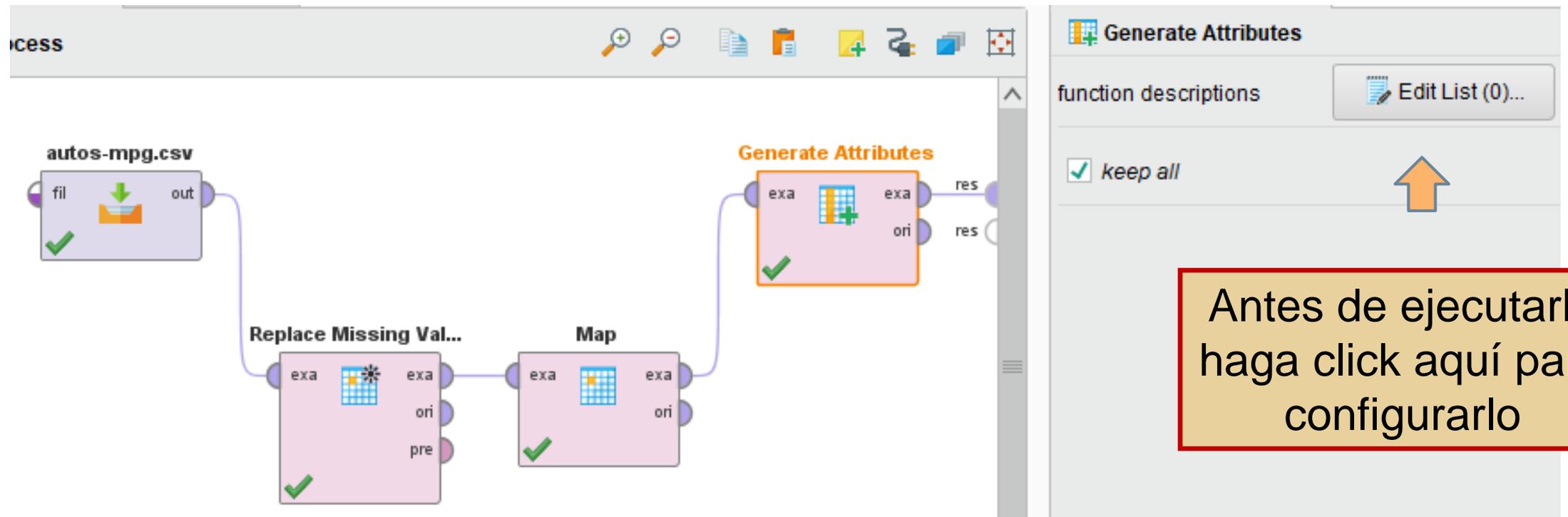
Generando un nuevo atributo



Generemos un nuevo atributo
utilizando el componente
Generate Attributes

Generando un nuevo atributo

□ Operador **Generate Attributes**



The image shows a data processing workflow in a software interface. The workflow consists of four operators connected in sequence: 1. A file reader operator labeled 'autos-mpg.csv' with 'fil' and 'out' ports. 2. A 'Replace Missing Val...' operator with 'exa', 'ori', and 'pre' ports. 3. A 'Map' operator with 'exa' and 'ori' ports. 4. A 'Generate Attributes' operator, which is highlighted with an orange border and has 'exa', 'ori', and 'res' ports. The 'Generate Attributes' operator is the focus of the right-hand panel.

The right-hand panel, titled 'Generate Attributes', shows the configuration for this operator. It includes a 'function descriptions' section with an 'Edit List (0)...' button. Below this, there is a checked checkbox labeled 'keep all'. An orange arrow points from this checkbox to a text box.

Antes de ejecutarlo haga click aquí para configurarlo

Generación de un nuevo atributo

Edit Parameter List: function descriptions

Edit Parameter List: function descriptions
List of functions to generate.

attribute name	function expressions
DECADA	if(model_year<80,"70s","80s")

Nombre del nuevo atributo

definición

Apply

Generación de un nuevo atributo

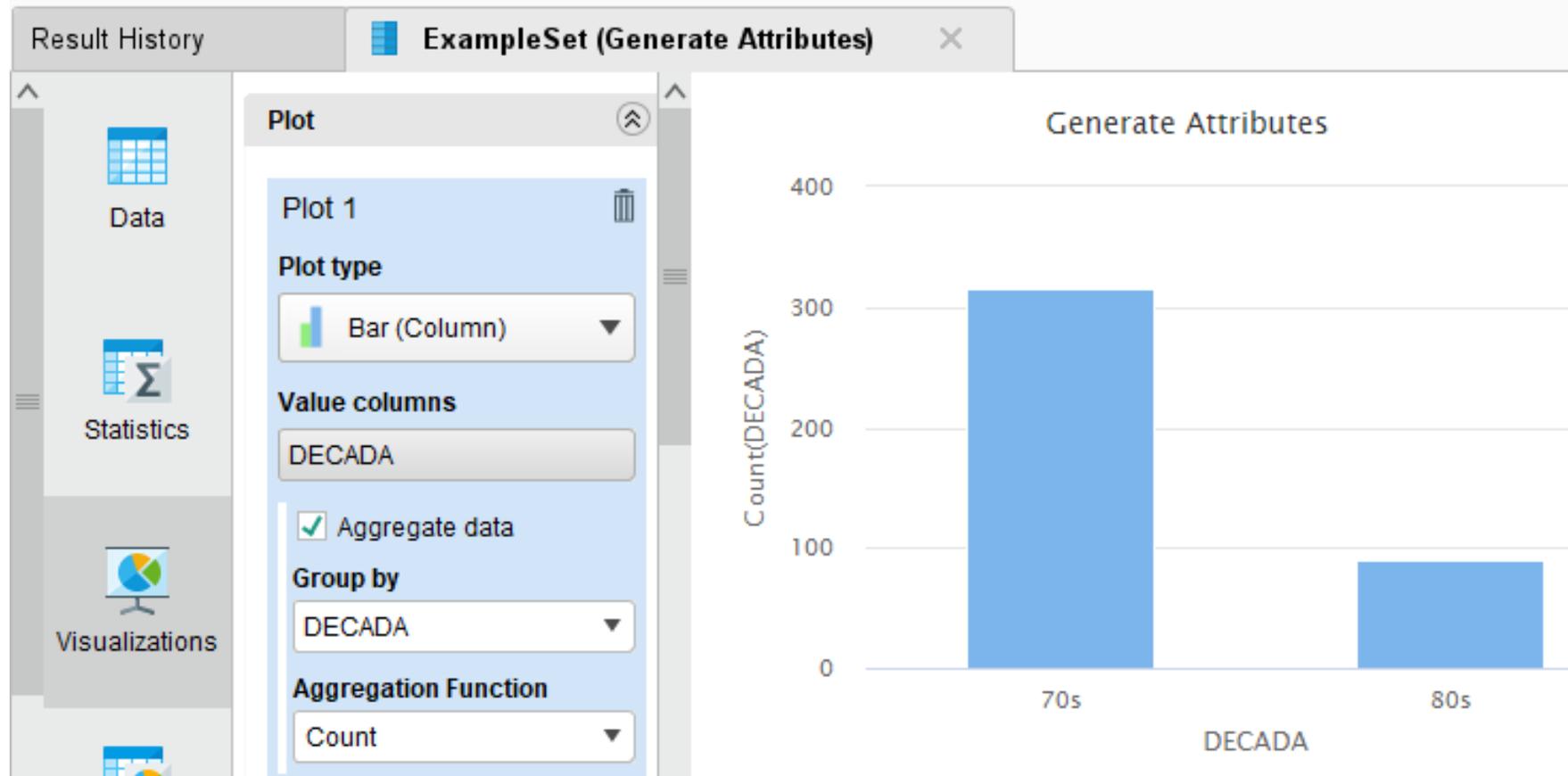
Edit Parameter List: function descriptions

Edit Parameter List: function descriptions
List of functions to generate.

attribute name	function expressions
DECADA	if(model_year<80,"70s","80s")

```
if (model_year<80 , "70s" , "80s")
```

Diagrama de barras del atributo generado



Transformación de atributos

□ DISCRETIZACION

- ▣ Algunos algoritmos de minería de datos sólo operan con atributos cualitativos. La discretización convierte los atributos numéricos en ordinales.

□ NUMERIZACION

- ▣ Es el proceso contrario a la discretización. Convierte atributos cualitativos en numéricos.

□ NORMALIZACION

- ▣ Permite expresar los valores de los atributos sin utilizar las unidades de medida originales facilitando su comparación y uso conjunto.

Discretización

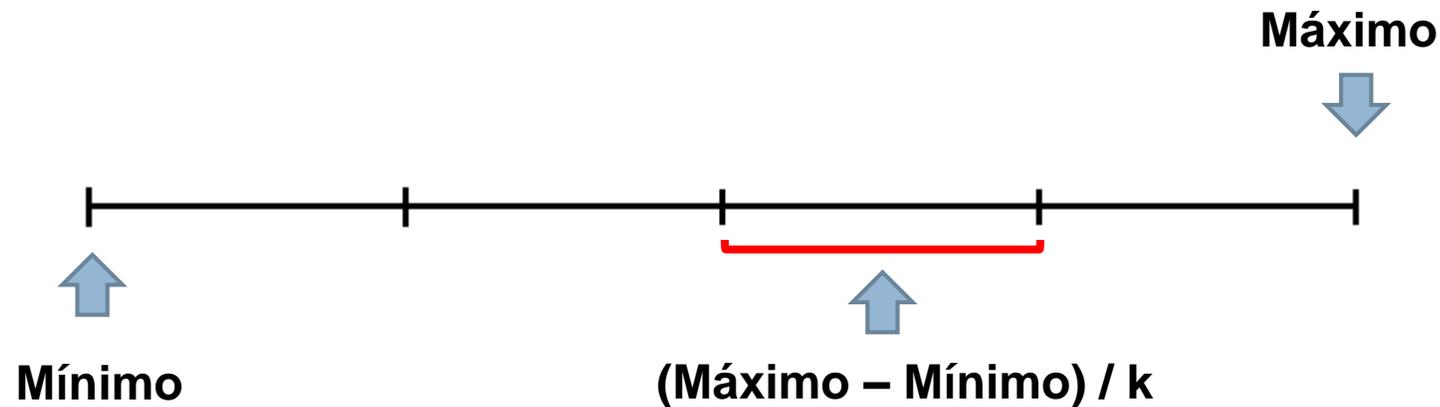
- Convierte un valor numérico en un nominal ordenado (que representa un intervalo o "*bin*")
- **Ejemplo:** Podemos transformar
 - ▣ la edad de la persona en categorías: $[0,12]$ niño, $(12-21)$ joven, $[21,65]$ adulto y >65 anciano.
 - ▣ La calificación de un alumno en: $[4,10]$ aprobado o $[0,4)$ desaprobado

Discretización

- Puede discretizarse en un número fijo de intervalos. El ancho del intervalo se calcula
 - ▣ Dividiendo el rango en partes iguales
 - ▣ Dividiendo la cantidad de ejemplos en partes iguales (igual frecuencia)
 - ▣ Indicando los límites de cada intervalo en forma manual.
- Averigüe por otras variantes de discretización

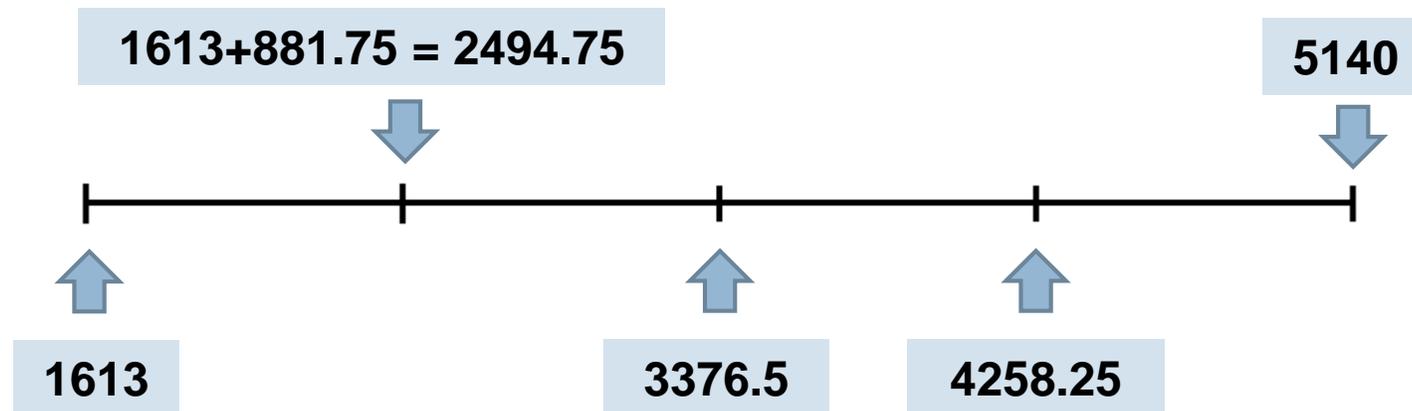
Discretización por rango

- El objetivo es dividir el rango del atributo (intervalo entre el máximo y el mínimo) en una cierta cantidad k de partes iguales.
- Los valores comprendidos en una misma parte serán asociados al mismo valor ordinal.
- Ejemplo: $k=4$



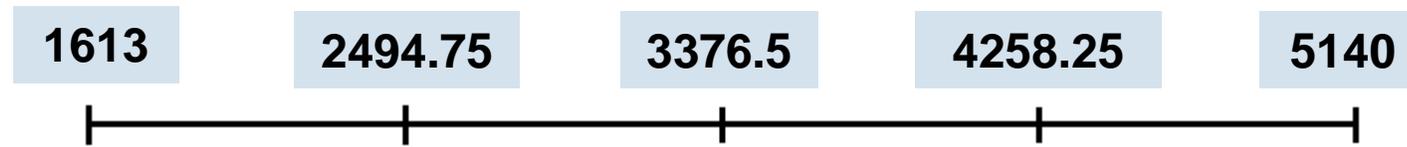
Discretización por rango

- **Ejemplo: Discretizar el atributo WEIGHT en 4 intervalos de igual longitud**
 - WEIGHT toma valores entre 1613 y 5140 libras. Si dividimos el rango en 4 partes iguales, cada una tendría una longitud de $(5140-1613)/4 = 881,75$



Discretización por rango

- **Ejemplo: Discretizar el atributo WEIGHT en 4 intervalos de igual longitud**



Valor	Intervalo	Frecuencia
range1	$(-\infty - 2494.750]$	147
range2	$(2494.750 - 3376.500]$	128
range3	$(3376.500 - 4258.250]$	90
range4	$(4258.250 - \infty]$	41

Discretización por rango

- WEIGHT discretizado en 4 intervalos de igual longitud

The image shows a data processing workflow. On the left, a 'mpg-autos.csv' file is loaded into a 'Discretize' operator. The operator has input ports 'fil' and 'out', and output ports 'exa', 'ori', and 'pre'. A blue arrow points from the operator to the configuration panel on the right. The configuration panel is titled 'Discretize (Discretize by Binning)' and includes the following settings:

- create view
- attribute filter type: single
- attribute: weight
- invert selection
- include special attributes
- number of bins: 4

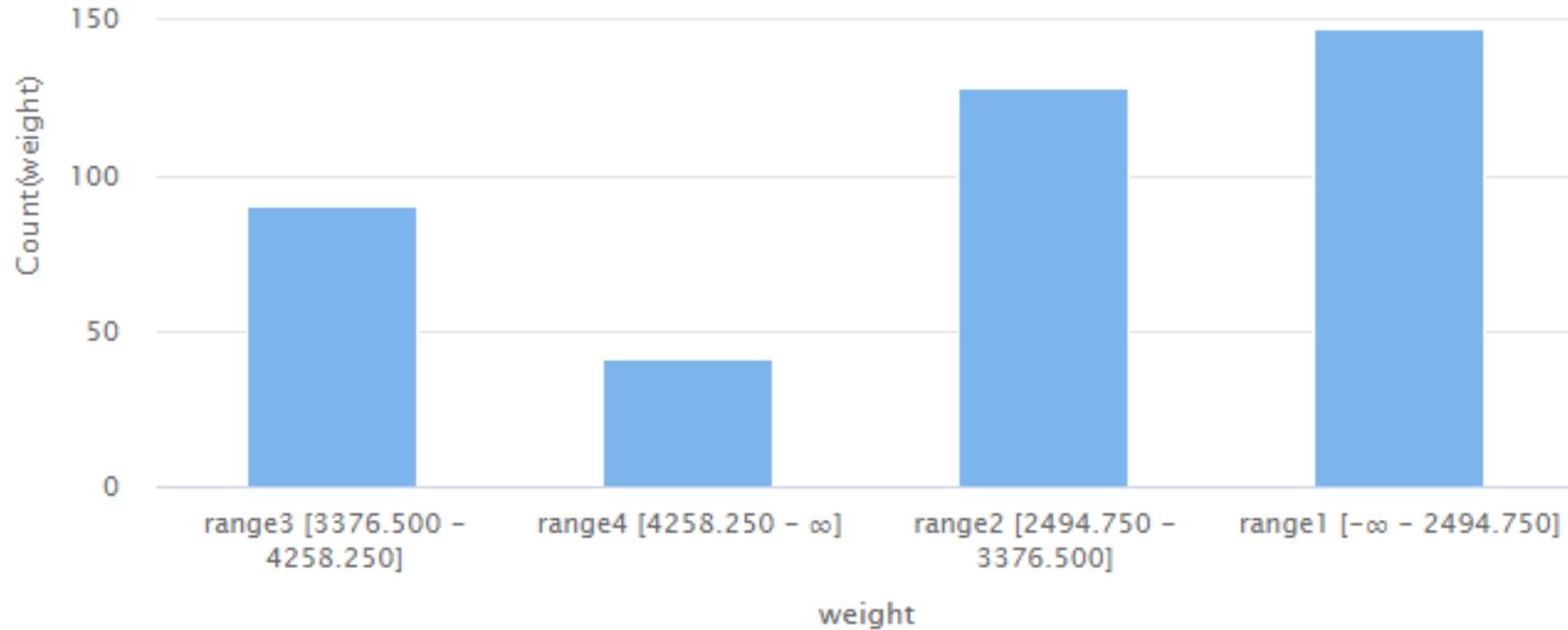
A red arrow points to the 'number of bins' field, which is highlighted by a red box.

Vamos a discretizar el atributo **WEIGHT** utilizando el operador **Discretize by Binning**

Indicar la cantidad de intervalos

Discretización por rango

- WEIGHT discretizado en 4 intervalos de igual longitud



Discretización por frecuencia

- El objetivo es dividir los valores del atributo numérico en k partes con la misma cantidad de valores en cada una de ellas.
- Nótese que el atributo debe tener al menos k valores diferentes.
- **Ejemplo: Discretizar WEIGHT en 4 intervalos de igual frecuencia**

Valor	Intervalo	Frecuencia
range1	$(-\infty - 2224.50]$	101
range2	$(2224.50 - 2822.50]$	102
range3	$(2822.50 - 3616.50]$	101
range4	$(3616.50 - \infty]$	102

Discretización por frecuencia

- WEIGHT discretizado en 4 intervalos de igual frecuencia

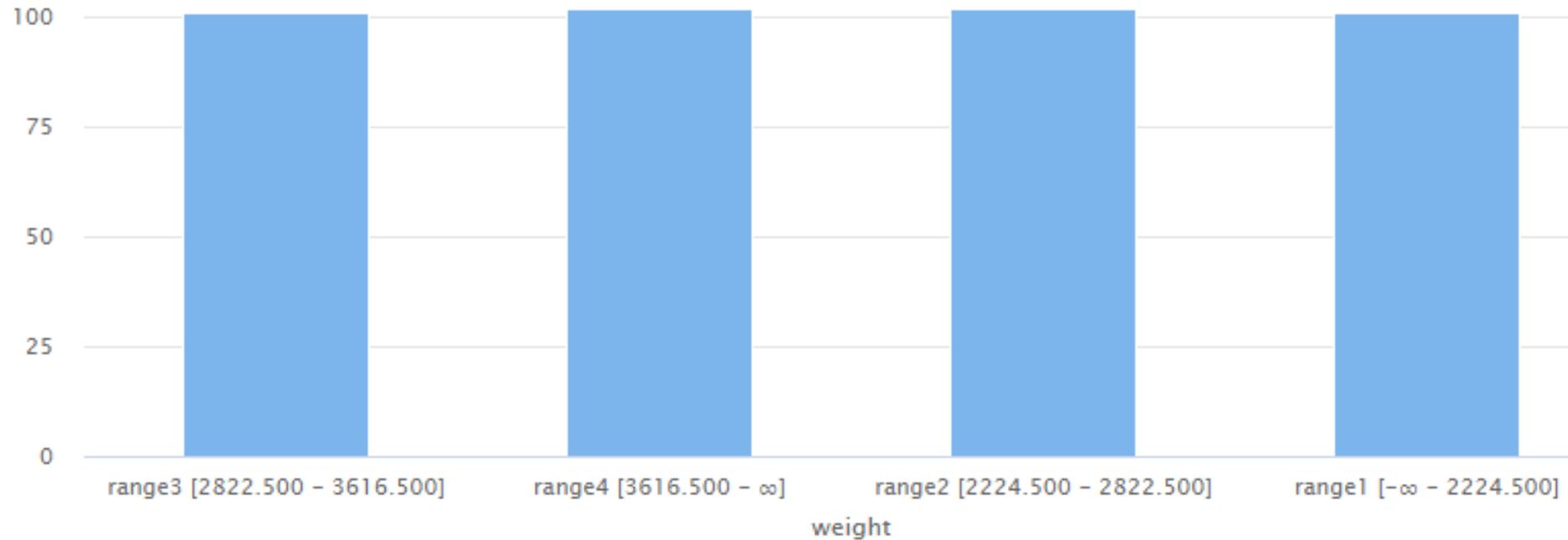
The image shows a data processing workflow. On the left, a 'Premios.csv' file is loaded into a 'Discretize' operator. The operator has 'exa' ports on the top and 'ori' and 'pre' ports on the right. Below the operator is a blue box labeled 'Operador Discretize by Frequency' with an arrow pointing up to the operator. To the right is a configuration panel for 'Discretize (Discretize by Frequency)'. The panel includes a 'create view' checkbox, 'attribute filter type' set to 'single', 'attribute' set to 'weight', 'invert selection' checkbox, 'include special attributes' checkbox, 'use sqrt of examples' checkbox, and 'number of bins' set to '4'. A red arrow points to the '4' in the 'number of bins' field, which is enclosed in a red box. A text box next to it says 'Indicar la cantidad de intervalos'.

Operador
Discretize by Frequency

Indicar la cantidad de intervalos

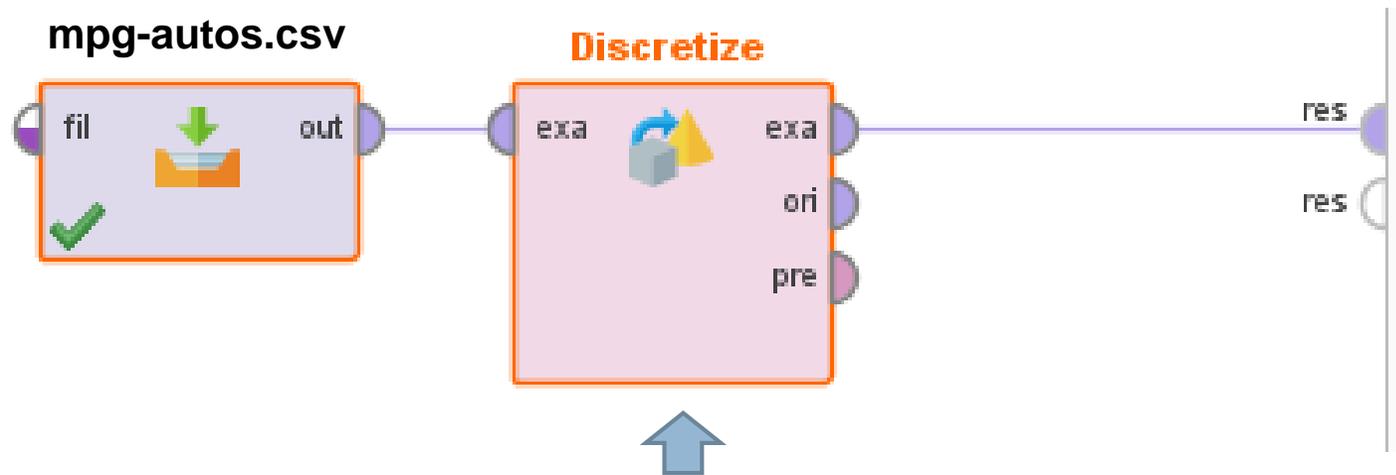
Discretización por frecuencia

- WEIGHT discretizado en 4 intervalos de igual frecuencia



Discretización especificada por el usuario

- Se indican los umbrales a utilizar en forma manual

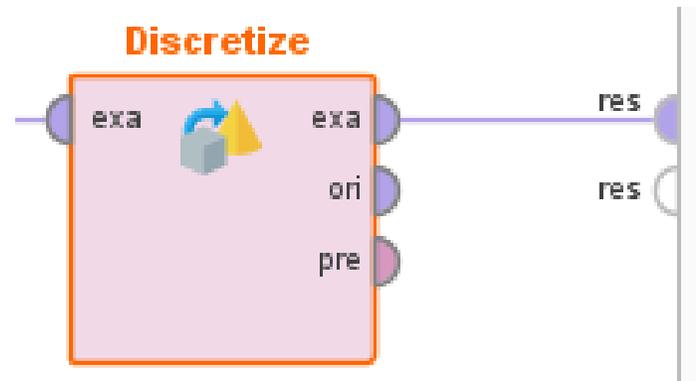


Vamos a discretizar el atributo **WEIGHT**
utilizando el operador

Discretize by User Specification

Discretización especificada por el usuario

Operador **Discretize by User Specification**



Se selecciona el atributo
WEIGHT

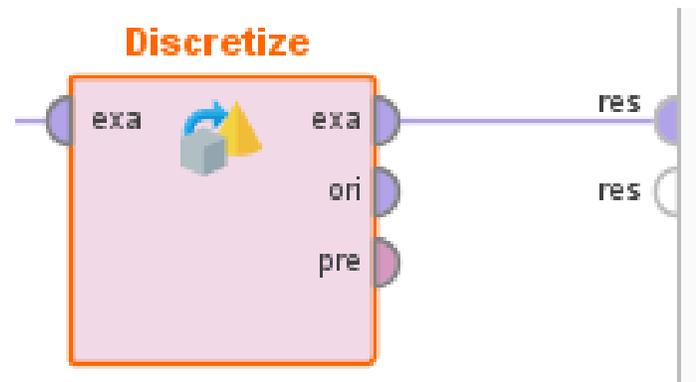
The screenshot shows the "Parameters" dialog for the "Discretize (Discretize by User Specification)" operator. The dialog has a close button (X) in the top right corner. The main title is "Discretize (Discretize by User Specification)". Below the title, there are several configuration options:

- create view
- attribute filter type: single
- attribute: weight
- invert selection
- include special attributes
- classes: Edit List (2)...

The "attribute filter type" and "attribute" fields are highlighted with a red border. Each field has an information icon (i) to its right.

Discretización especificada por el usuario

□ Operador **Discretize by User Specification**



Aquí se indican los intervalos

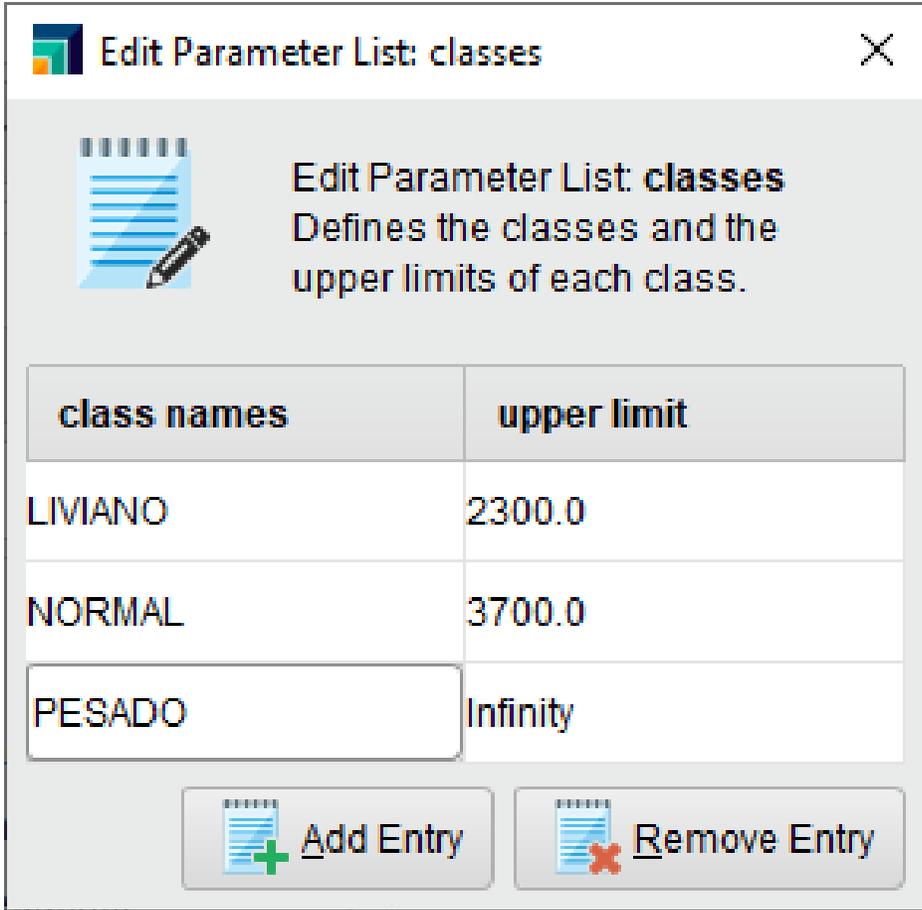
The screenshot shows a "Parameters" dialog box for the "Discretize (Discretize by User Specification)" operator. The dialog has a close button (X) in the top right corner. The parameters are as follows:

- create view
- attribute filter type: single
- attribute: weight
- invert selection
- include special attributes
- classes: Edit List (2)...

The "classes" parameter and its "Edit List (2)..." button are highlighted with a red rectangular box. A blue arrow points from the text box "Aquí se indican los intervalos" to this red box.

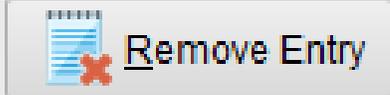
Discretización especificada por el usuario

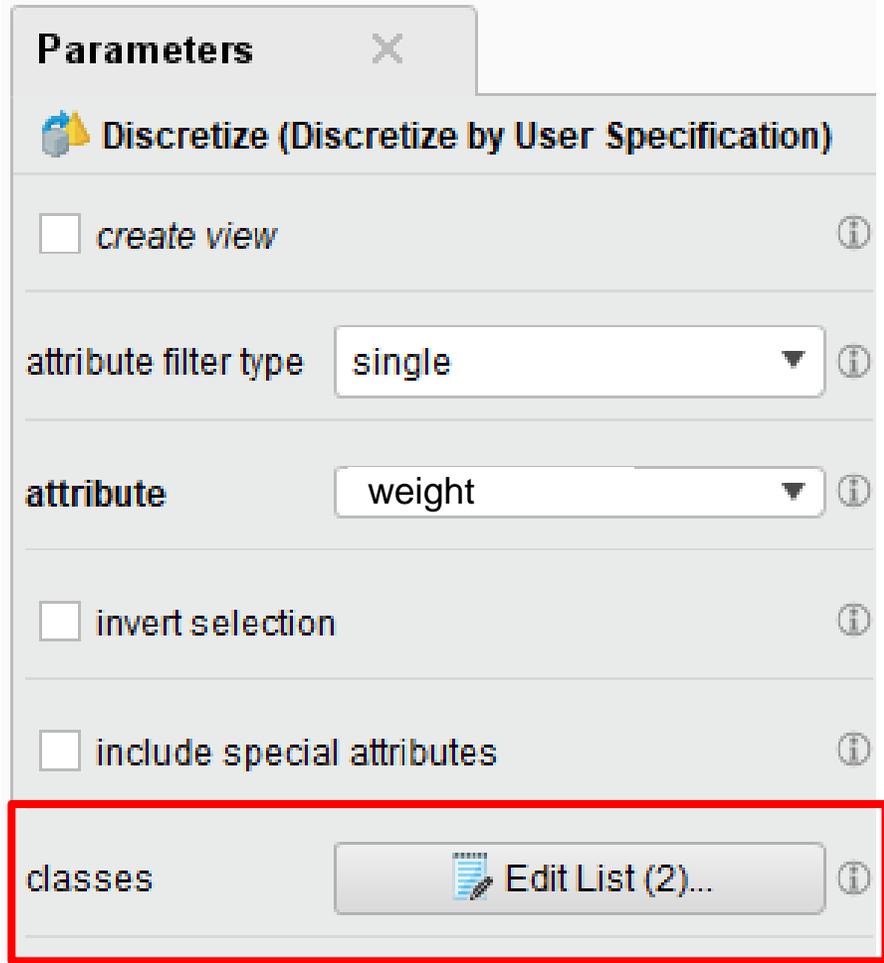
Operador **Discretize by User Specification**

 Edit Parameter List: classes

Defines the classes and the upper limits of each class.

class names	upper limit
LIVIANO	2300.0
NORMAL	3700.0
PESADO	Infinity

 Add Entry  Remove Entry

 Parameters

 Discretize (Discretize by User Specification)

create view

attribute filter type: single

attribute: weight

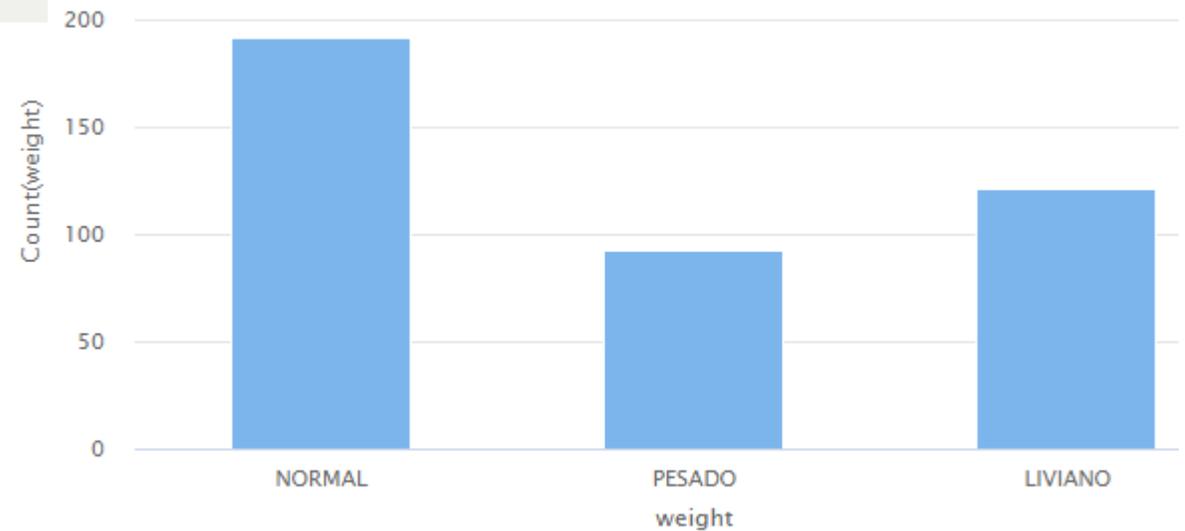
invert selection

include special attributes

classes 

Discretización especificada por el usuario

Valor	Intervalo	Frecuencia
LIVIANO	$(-\infty - 2300]$	147
NORMAL	$(2300 - 3700]$	128
PESADO	$(3700 - \infty)$	90

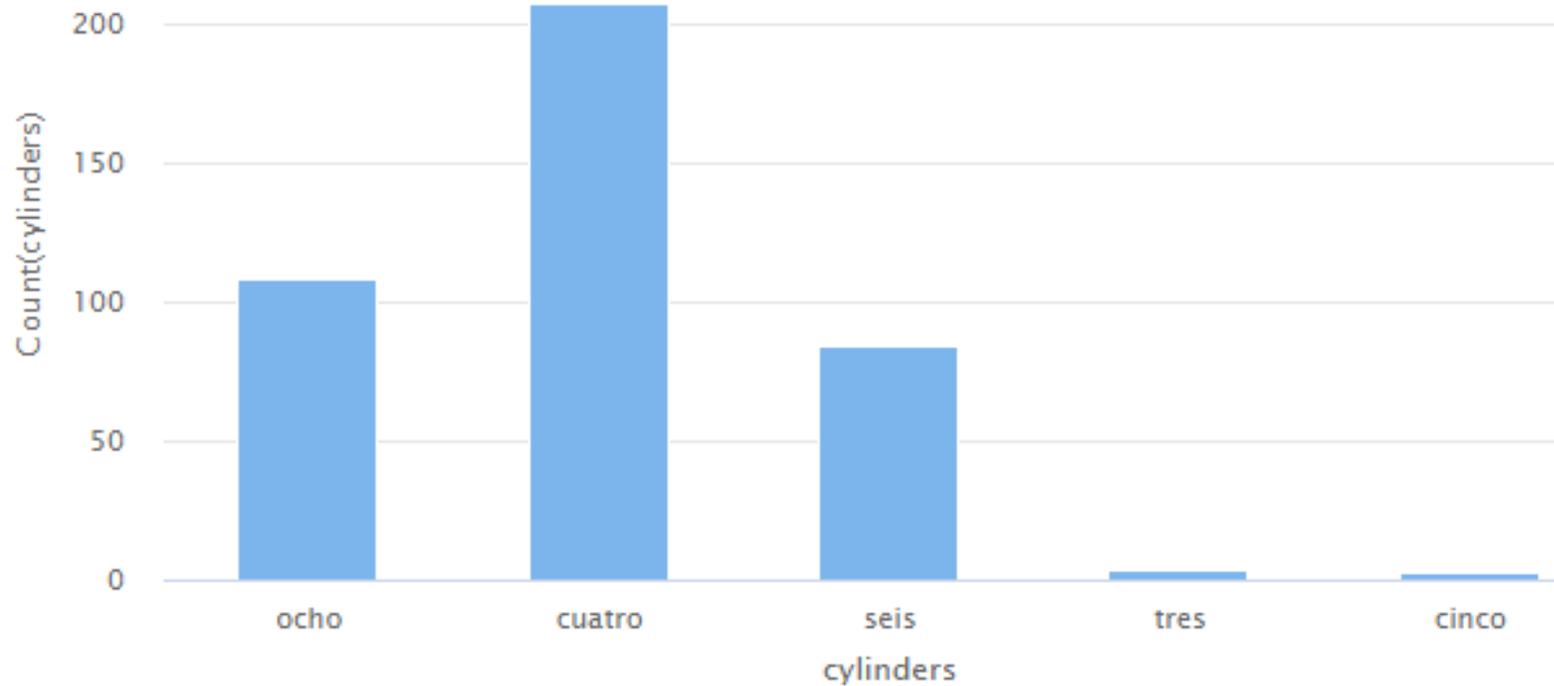


Numerización

- En ocasiones los atributos nominales u ordinales deben convertirse en números.
- Para los nominales suele utilizarse una representación binaria y para los ordinales suele utilizarse una representación entera.
- Es importante considerar que si se numeran en forma correlativa los valores de un atributo nominal se agrega un orden que originalmente no está presente en la información disponible.

Numerizando un atributo ordinal

- Asignaremos a cada valor del atributo ordinal CYLINDERS su correspondiente valor numérico usando el operador **MAP**

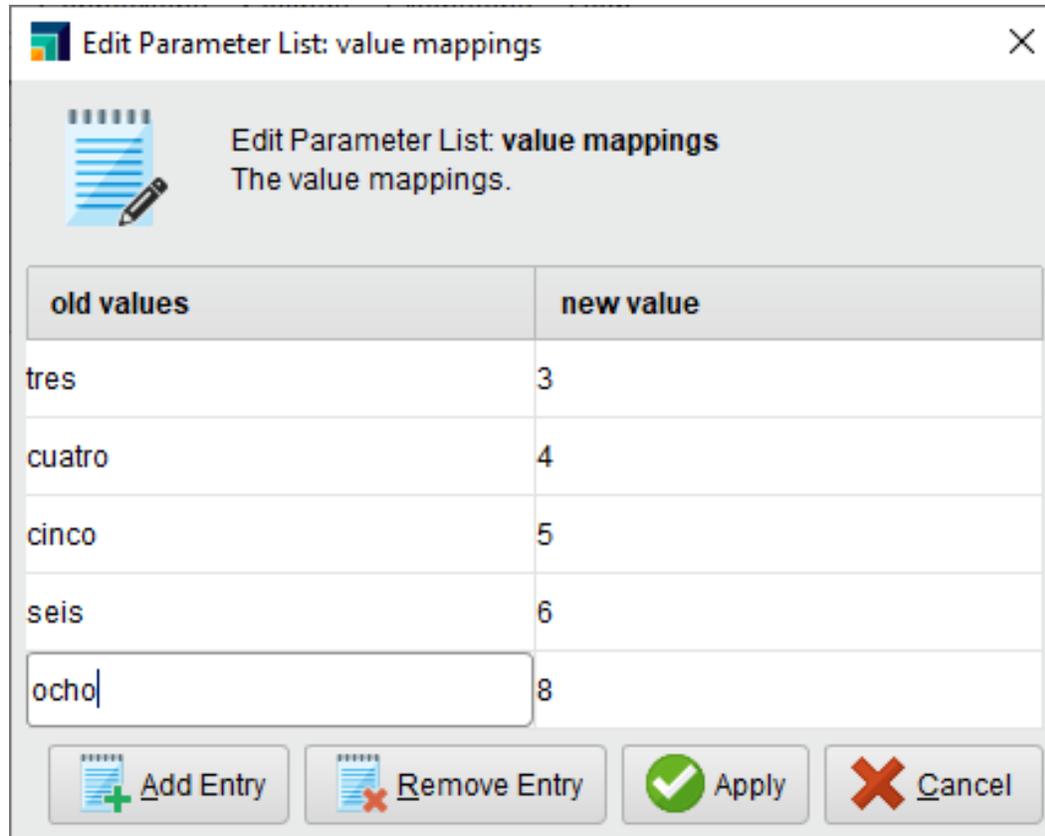


Numerizando el atributo CYLINDERS

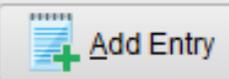
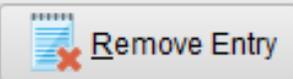
The image shows a data processing workflow in a software interface. On the left, a file named 'autos-mpg.csv' is loaded into a 'Map' operation. The 'Map' operation is connected to a 'res' output. The right-hand side of the image shows the configuration panel for the 'Map' operation, which includes the following settings:

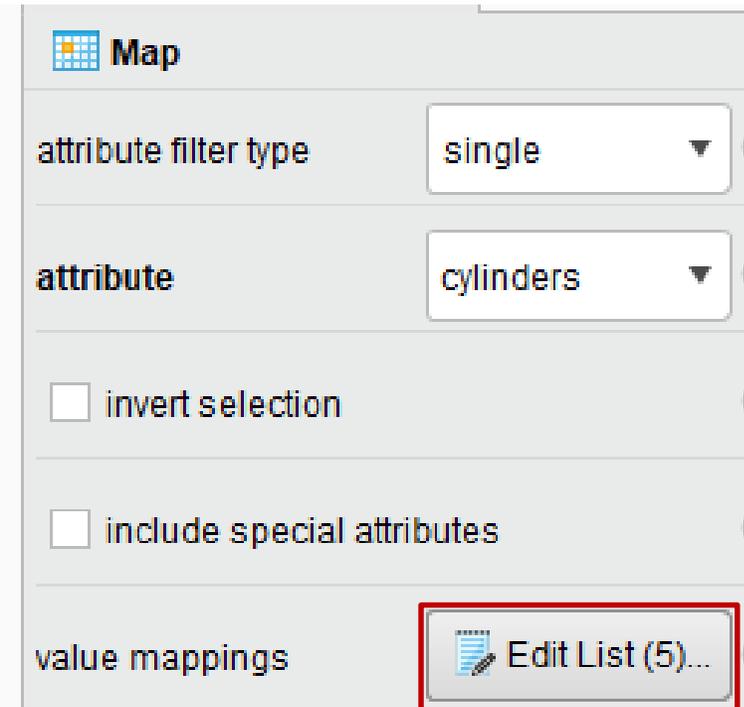
- attribute filter type:** single
- attribute:** cylinders
- invert selection
- include special attributes
- value mappings:** Edit List (5)...

Numerizando el atributo CYLINDERS

The dialog box is titled "Edit Parameter List: value mappings" and contains a table with two columns: "old values" and "new value". The table lists the following mappings: "tres" to "3", "cuatro" to "4", "cinco" to "5", "seis" to "6", and "ocho" to "8". Below the table are four buttons: "Add Entry", "Remove Entry", "Apply", and "Cancel".

old values	new value
tres	3
cuatro	4
cinco	5
seis	6
ocho	8

The "Map" configuration panel shows the following settings: "attribute filter type" is set to "single", "attribute" is set to "cylinders", "invert selection" is unchecked, and "include special attributes" is unchecked. The "value mappings" section has a button labeled "Edit List (5)..." which is highlighted with a red box.

Map

attribute filter type: single

attribute: cylinders

invert selection

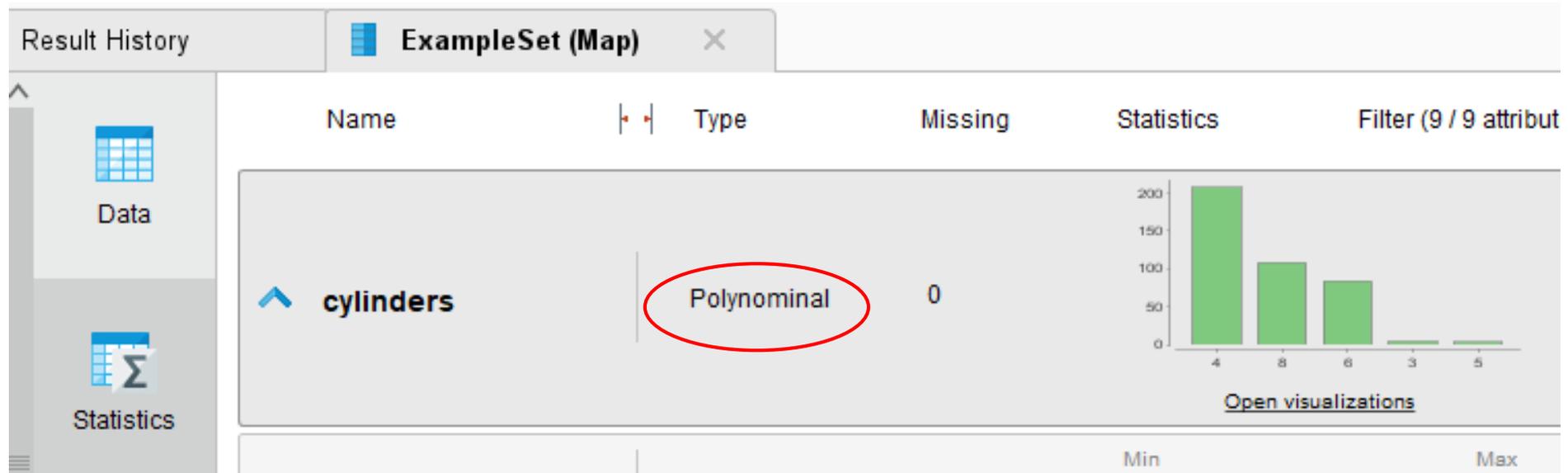
include special attributes

value mappings: 



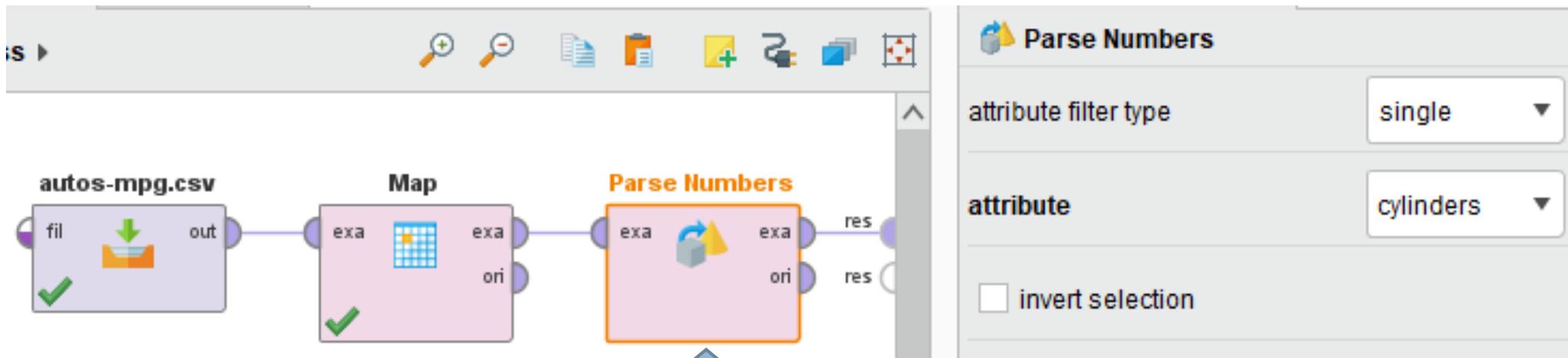
Numerizando el atributo CYLINDERS

- Ejecute y verifique que el atributo continua siendo cualitativo



Numerizando el atributo CYLINDERS

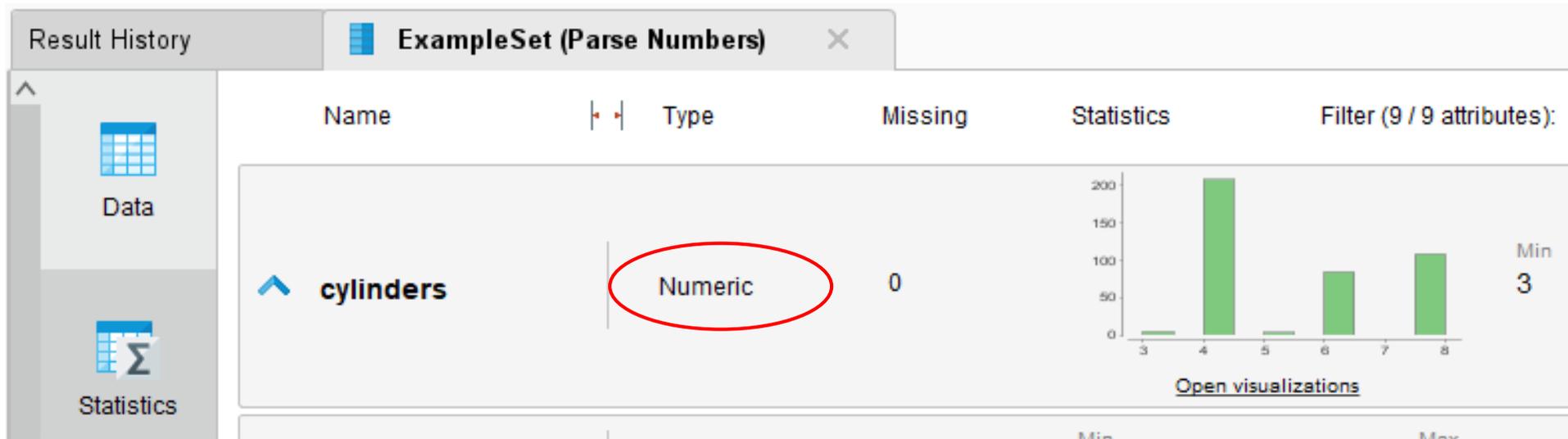
□ Operador **Parse Numbers**



Cambia el tipo de dato del atributo a numérico

Numerizando el atributo CYLINDERS

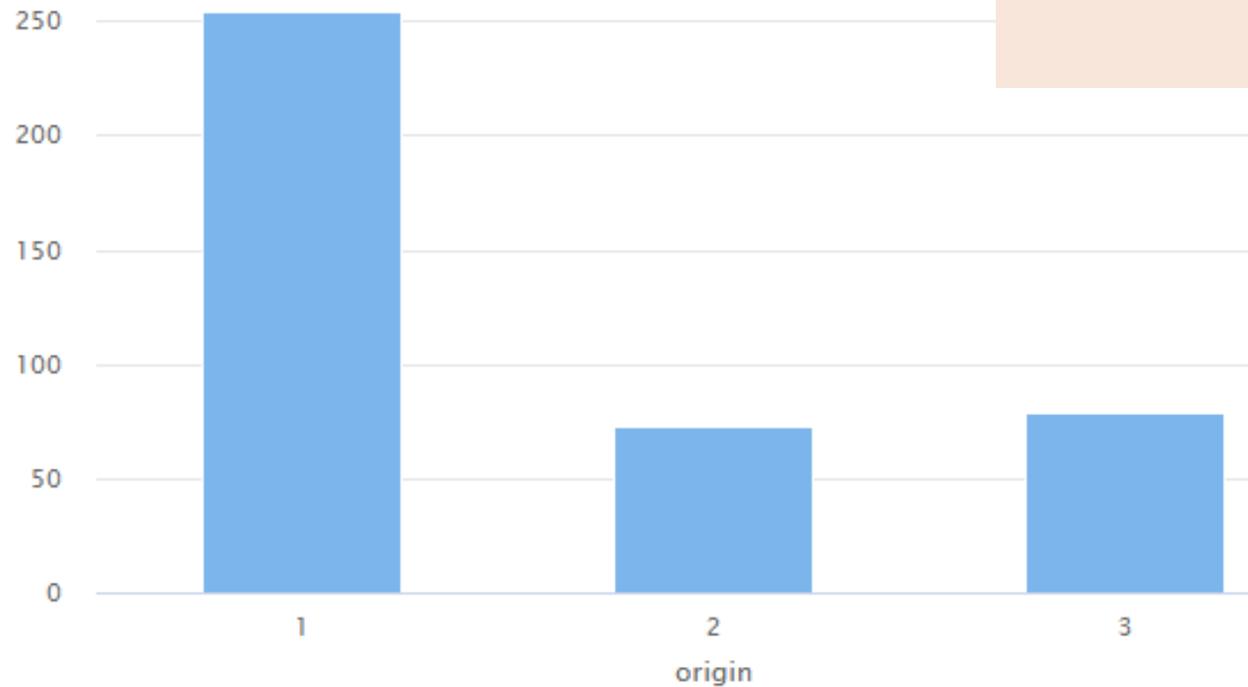
- Ejecute y verifique que el atributo es numérico



Numerización Binaria (dummy)

- La numerización binaria reemplaza al atributo nominal por tantos atributos numéricos binarios como valores distintos pueda tomar.
- Las denominaciones de estos nuevos atributos surgen de igualar el nombre original con cada uno de los posibles valores.
- Para un mismo ejemplo sólo uno de estos nuevos atributos tendrá valor 1 y el resto 0.

Atributo ORIGIN



Si bien ORIGIN contiene números, no se trata de un atributo cuantitativo ¿ Por qué ?

- 1 - USA
- 2 - Europe
- 3 - Japan

Read CSV – data set meta data information

The image shows a software interface for reading a CSV file. On the left, a workflow diagram shows a 'Read CSV' component connected to a 'Replace Missing Values' component. The 'Read CSV' component has inputs 'fil' and 'out', and a green checkmark. The 'Replace Missing Values' component has inputs 'exa' and 'ori', and outputs 'exa' and 'pre', with a green checkmark. On the right, a 'Parameters' panel for 'autos-mpg.csv (Read CSV)' is shown. It includes settings for 'date format', 'first row as names' (checked), 'time zone' (SYSTEM), 'locale' (English (United...)), 'encoding' (windows-1252), and 'read all values as polynomial' (unchecked). At the bottom, the 'data set meta data information' section has an 'Edit List (9)...' button highlighted with a red box. A yellow arrow points from the text box below to this button.

autos-mpg.csv (Read CSV)

date format ⓘ

first row as names ⓘ

time zone ⓘ

locale ⓘ

encoding ⓘ

read all values as polynomial ⓘ

data set meta data information ⓘ

Permite seleccionar atributos y modificar su tipo y role

Read CSV – data set meta data information

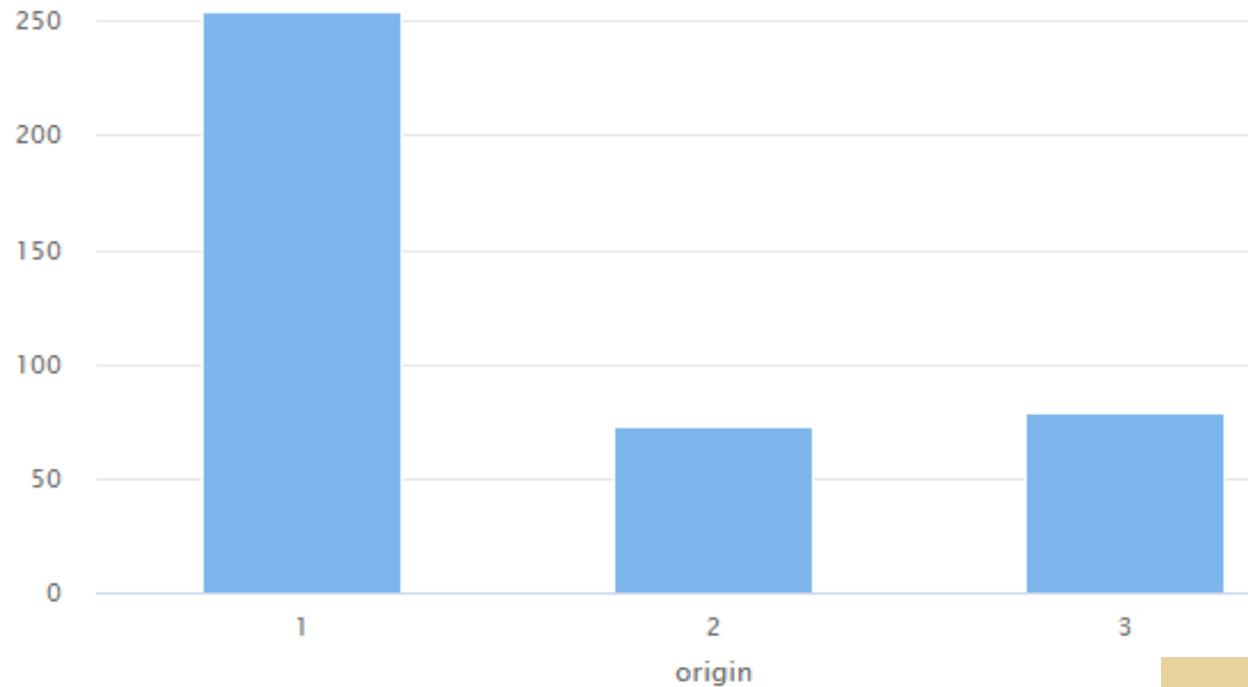


Edit Parameter List: data set meta data information

Edit Parameter List: data set meta data information
The meta data information

column...	attribute meta data information
0	mpg <input checked="" type="checkbox"/> column selected real attribute
1	cylinders <input checked="" type="checkbox"/> column selected polynomial attribute
2	displacement <input checked="" type="checkbox"/> column selected integer attribute
3	horsepower <input checked="" type="checkbox"/> column selected real attribute
4	weight <input checked="" type="checkbox"/> column selected integer attribute
5	acceleration <input checked="" type="checkbox"/> column selected integer attribute
6	model_year <input checked="" type="checkbox"/> column selected integer attribute
7	origin <input checked="" type="checkbox"/> column selected polynomial attribute
8	car_name <input checked="" type="checkbox"/> column selected polynomial attribute

Atributo ORIGIN



1 - USA
2 - Europe
3 - Japan

Puede usarse el operador **MAP** para asignar las nuevas etiquetas

Atributo ORIGIN

The image shows a data processing workflow in a software interface. The workflow consists of three main components connected in a sequence:

- autos-mpg.csv**: A file input tool with ports labeled 'fil' and 'out'.
- Replace Missing Val...**: A tool with an 'exa' input port and 'ori' and 'pre' output ports.
- Map**: A tool with 'exa' input and output ports.

The 'Map' tool is currently selected, and its configuration panel is open on the right. The panel is titled 'Parameters' and 'Map'. A red arrow points to the 'Map' title. The configuration panel includes the following settings:

- attribute filter type**: set to 'single'.
- attribute**: set to 'origin'.
- invert selection**
- include special attributes**
- value mappings**: with an 'Edit List (3)...' button.

Atributo GENRE1

The image shows a data processing workflow in a software interface. The workflow consists of three steps:

- autos-mpg.csv**: A file input step with ports labeled 'fil' and 'out'.
- Replace Missing Val...**: A data cleaning step with ports labeled 'exa', 'ori', and 'pre'.
- Map**: An attribute filtering step with ports labeled 'exa' and 'ori'.

The **Map** step is selected, and its parameters are shown in the **Parameters** panel on the right:

- Map**: The step name.
- attribute filter type**: Set to **single**.
- attribute**: Set to **origin**.
- invert selection**
- include special attributes**
- value mappings**: Includes an **Edit List (3)...** button, which is highlighted with a red box and a yellow arrow pointing to it.

Atributo GENRE1

The image shows a QGIS interface with two main windows. On the left, a file browser window titled 'autos-mpg.csv' is open, showing a file icon with a green checkmark. In the center, the 'Edit Parameter List: value mappings' dialog is displayed. It has a title bar with a close button and a toolbar with icons for zooming and editing. The dialog contains a table with two columns: 'old values' and 'new value'. The table has three rows with values 1, 2, and 3 in the first column, and USA, Europe, and Japan in the second column. Below the table are three buttons: 'Add Entry', 'Remove Entry', and 'Apply'. On the right, the 'Parameters' panel is visible, showing a 'Map' tab. It has several settings: 'attribute filter type' set to 'single', 'attribute' set to 'origin', and two checkboxes for 'invert selection' and 'include special attributes'. At the bottom, there is a 'value mappings' section with a button labeled 'Edit List (3)...' which is highlighted with a red rectangle. A yellow arrow points upwards towards this button.

Edit Parameter List: value mappings

Edit Parameter List: value mappings
The value mappings.

old values	new value
1	USA
2	Europe
3	Japan

Add Entry Remove Entry Apply

Parameters

Map

attribute filter type single

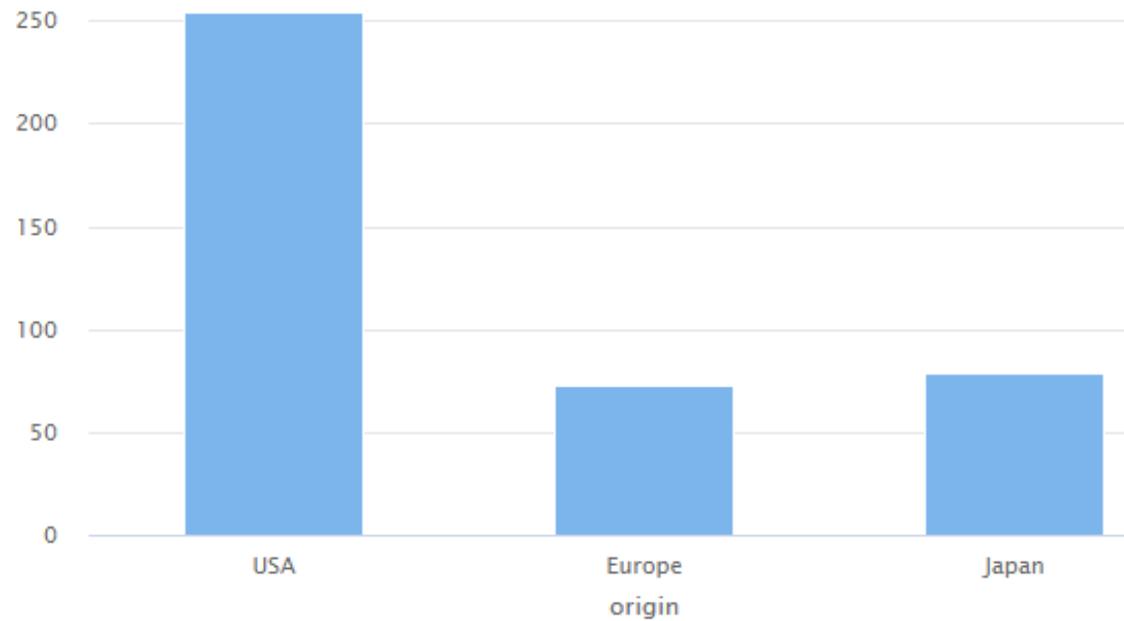
attribute origin

invert selection

include special attributes

value mappings Edit List (3)...

Atributo ORIGIN

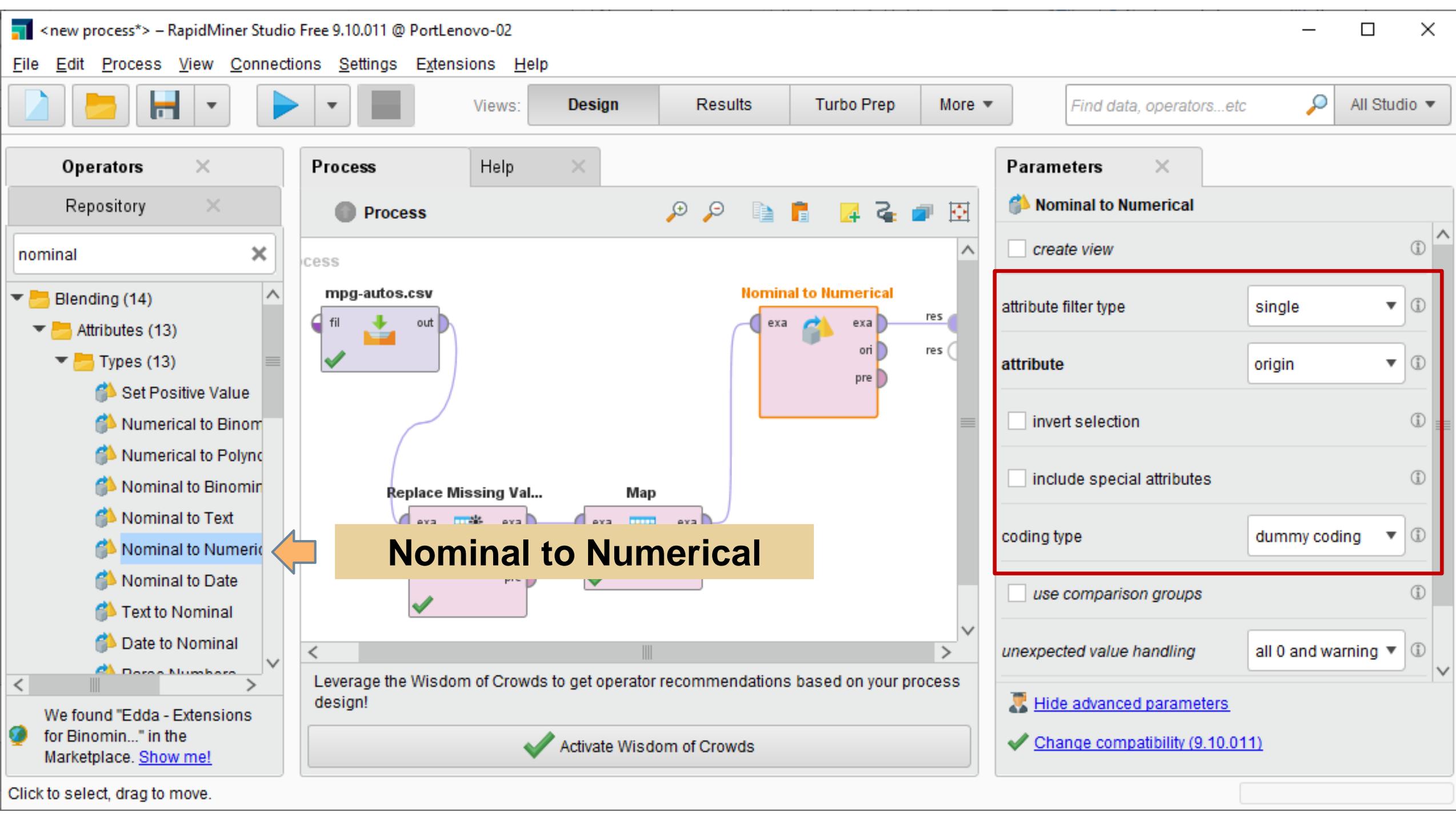


Nominal values



Index	Nominal value	Absolute count	Fraction
1	USA	254	0.626
2	Japan	79	0.195
3	Europe	73	0.180

Close



Numerización Binaria de ORIGIN

Row No.	origin = USA	origin = Europe	origin = Japan	mpg	cylinders	displacement	horsepower
21	0	0	1	24	cuatro	1130	95
22	1	0	0	22	seis	1980	95
23	1	0	0	18	seis	1990	97
24	1	0	0	21	seis	2000	85
25	0	0	1	27	cuatro	9700	88
26	0	1	0	26	cuatro	9700	46
27	0	1	0	25	cuatro	1100	87
28	0	1	0	24	cuatro	1070	90
29	0	1	0	25	cuatro	1040	95
30	0	1	0	26	cuatro	1210	113

Transformación de atributos

□ DISCRETIZACION

- ▣ Algunos algoritmos de minería de datos sólo operan con atributos cualitativos. La discretización convierte los atributos numéricos en ordinales.

□ NUMERIZACION

- ▣ Es el proceso contrario a la discretización. Convierte atributos cualitativos en numéricos.

□ NORMALIZACION

- ▣ Permite expresar los valores de los atributos sin utilizar las unidades de medida originales facilitando su comparación y uso conjunto.

Normalización

- Se aplica según el modelo que se va a construir.
- La más común es la **normalización lineal uniforme**

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

- Es muy sensible a valores fuera de rango (outliers).

Normalización

- Existen otras transformaciones. Por ejemplo, si los datos tienen distribución normal se pueden **tipificar**

$$X' = \frac{X - \text{media}(X)}{\text{desviacion}(X)}$$

- De esta forma los datos se distribuyen normalmente alrededor de 0 con desviación 1.

Normalización – Operador Normalize

The image shows a data processing workflow in a visual programming environment. The workflow consists of three main operators: 'mpg-autos.csv', 'Replace Missing Val...', and 'Normalize'. The 'mpg-autos.csv' operator is connected to 'Replace Missing Val...', which is connected to 'Select Attributes', which is finally connected to 'Normalize'. The 'Normalize' operator is highlighted with an orange border. Below the 'Select Attributes' operator, a configuration panel is shown with 'attribute filter type' set to 'value_type' and 'value type' set to 'numeric'. To the right of the 'Normalize' operator, another configuration panel is shown with 'attribute filter type' set to 'all' and 'method' set to 'range transformation'. A dropdown menu is open for the 'method' property, showing options: 'range transformation', 'Z-transformation', 'range transformation', 'proportion transformation', and 'interquartile range'. The 'range transformation' option is highlighted in blue.

mpg-autos.csv

Replace Missing Val...

Select Attributes

Normalize

Select Attributes

attribute filter type: value_type

value type: numeric

Normalize

create view:

attribute filter type: all

invert selection:

include special attributes:

method: range transformation

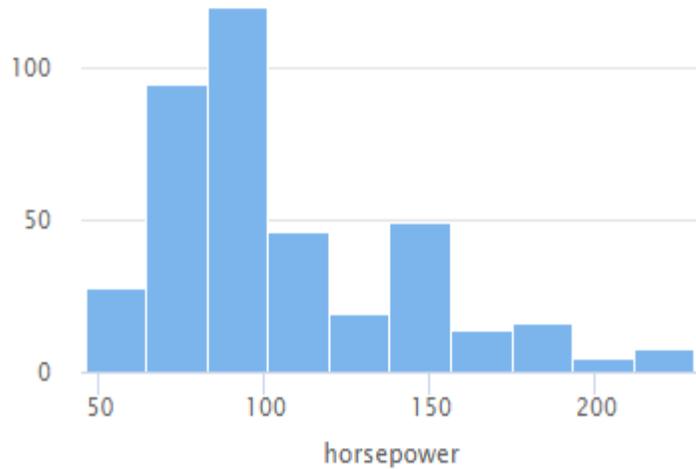
min: range transformation

interquartile range

Para seleccionar sólo los atributos numéricos

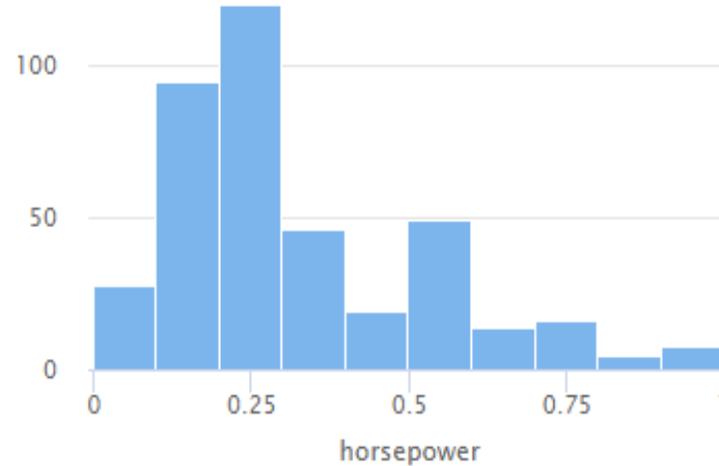
Normalización del atributo HORSEPOWER

Original



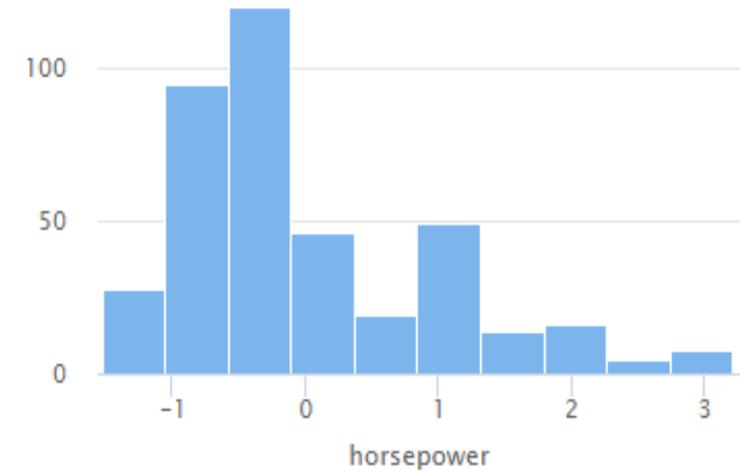
X

Lineal
(range transformation)



$$X' = \frac{X - \min(X)}{\max(X) - \min(X)}$$

Tipificación
(Z-transformation)



$$X' = \frac{X - \text{media}(X)}{\text{desvio}(X)}$$

Indique similitudes y diferencias entre los histogramas

Comparación de atributos numéricos

□ Valores originales

#	mpg	displacement	horsepower	weight	acceleration	model_year
1	18	3070	130	3504	120	70
2	15	3500	165	3693	115	70
3	18	3180	150	3436	110	70
...
123	15	3500	145	4082	130	73
124	16	4000	230	4278	950	73
125	29	6800	49	1867	195	73
...
404	32	1350	84	2295	116	82
405	28	1200	79	2625	186	82
406	31	1190	82	2720	194	82

Comparación de atributos numéricos

- Valores normalizados linealmente entre 0 y 1

#	mpg	displacement	horsepower	weight	acceleration	model_year
1	0.239	0.235	0.457	0.536	0.046	0
2	0.160	0.284	0.647	0.590	0.040	0
3	0.239	0.248	0.565	0.517	0.034	0
...
123	0.160	0.284	0.538	0.700	0.057	0.250
124	0.186	0.341	1.000	0.756	1.000	0.250
125	0.532	0.659	0.016	0.072	0.132	0.250
...
404	0.612	0.040	0.207	0.193	0.041	1
405	0.505	0.023	0.179	0.287	0.122	1
406	0.585	0.022	0.196	0.314	0.131	1

Comparación de atributos numéricos

- Valores normalizados utilizando media y desvío

#	mpg	displacement	horsepower	weight	acceleration	model_year
1	-0.713	-0.275	0.648	0.619	-0.772	-1.580
2	-1.100	-0.133	1.557	0.842	-0.876	-1.580
3	-0.713	-0.239	1.167	0.539	-0.979	-1.580
...
123	-1.100	-0.133	1.037	1.302	-0.565	-0.779
124	-0.971	0.032	3.246	1.533	16.414	-0.779
125	0.709	0.954	-1.457	-1.313	0.781	-0.779
...
404	1.097	-0.841	-0.548	-0.808	-0.855	1.622
405	0.580	-0.891	-0.678	-0.418	0.594	1.622
406	0.967	-0.894	-0.600	-0.306	0.760	1.622

Semillas de trigo

- El archivo **SEMILLAS.csv** contiene información de granos que pertenecen a tres variedades diferentes de trigo: Kama, Rosa y Canadiense.
 - ▣ área A ,
 - ▣ perímetro P ,
 - ▣ compacidad $C = 4 * \pi * A / P^2$,
 - ▣ longitud del núcleo,
 - ▣ ancho del núcleo,
 - ▣ coeficiente de asimetría
 - ▣ longitud del surco del núcleo

Analice estos datos y explique las relaciones encontradas

Resumen

PREPARACION DE LOS DATOS

- Detección de valores atípicos (diagramas de caja y bigotes)
- Completar datos faltantes
- Operador MAP
- Generación de características o atributos nuevos
- Transformaciones
 - ▣ Discretización por rango, por frecuencia e indicada por el usuario
 - ▣ Numerización: codificación entera y codificación binaria
 - ▣ Normalización: Lineal y Estandarización (o tipificación)