

Introduction to Data Science



Dr. José Ramón Iglesias

DSP-ASIC BUILDER GROUP
Director Semillero TRIAC
Ingenieria Electronica
Universidad Popular del Cesar



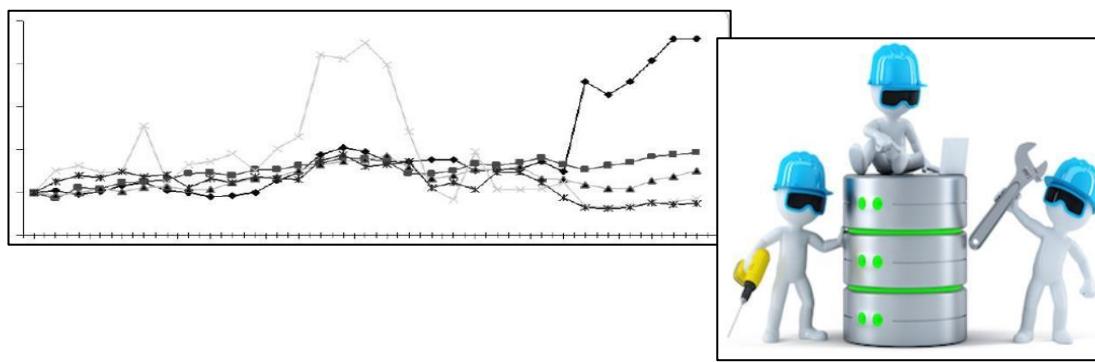
MINERÍA DE DATOS

Introducción

- Los avances tecnológicos hacen que las **capacidades para generar y almacenar datos** se incrementen día a día.
 - Automatización de todo tipo de transacciones
 - Comerciales, negocios, gubernamentales, científicas.
 - Avances en la recopilación de datos (Lectores)
 - Mejora en la relación precio-capacidad de los dispositivos de almacenamiento masivo.



Cómo se originan los datos?



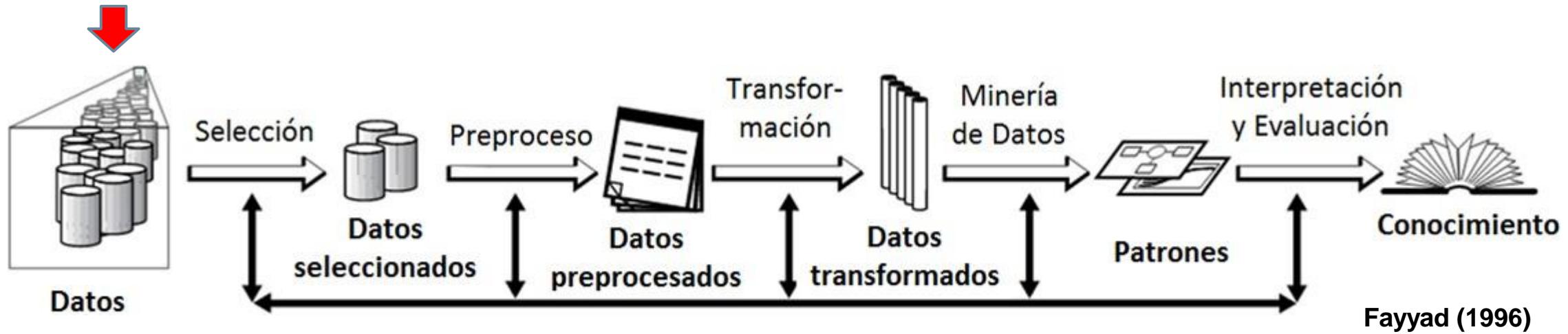
¿Qué es Minería de Datos?

- Es el área informática que busca descubrir **patrones** en grandes volúmenes de datos.

- Características
 - Válidos
 - Novedosos
 - Potencialmente útiles
 - Comprensibles

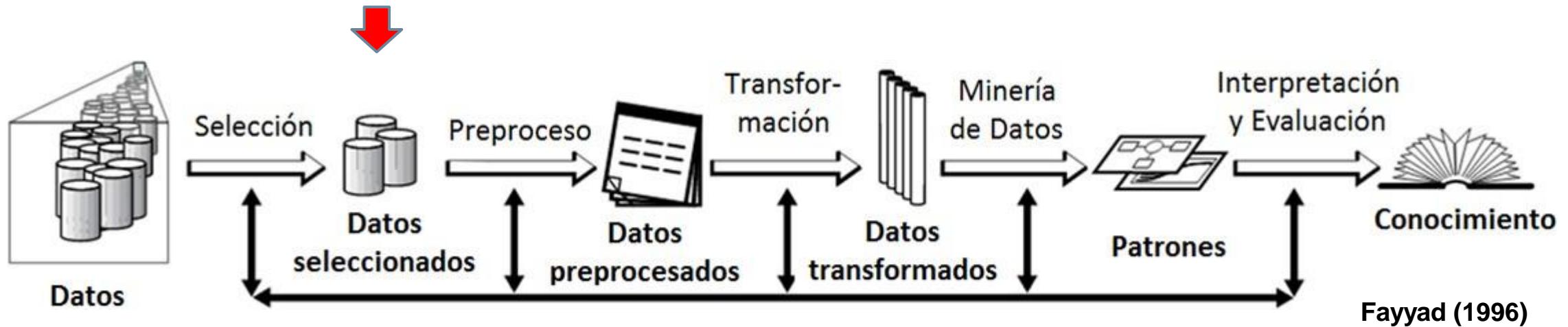


Extracción de Conocimiento (proceso KDD)



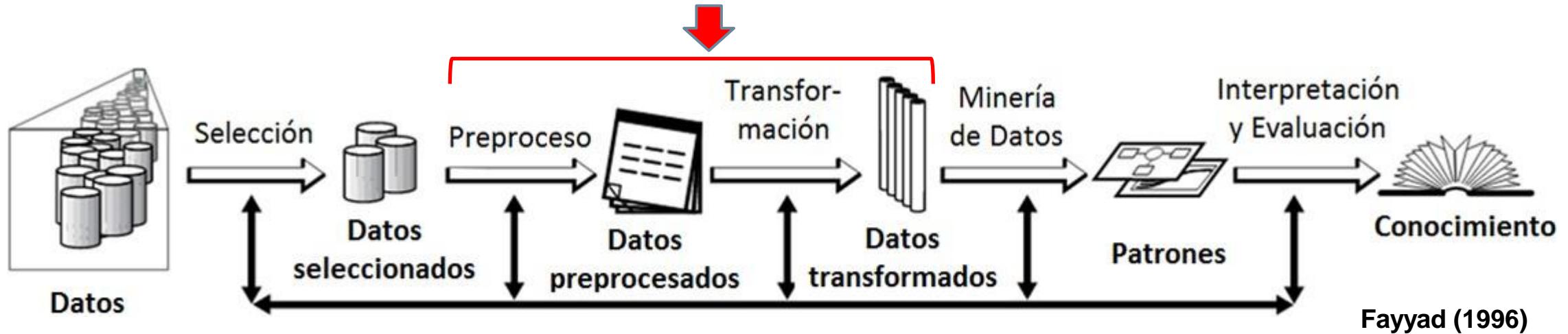
- Generalmente registrado en forma previa al proceso de KDD.
- Almacena información histórica
- No necesariamente centralizada

Extracción de Conocimiento (proceso KDD)



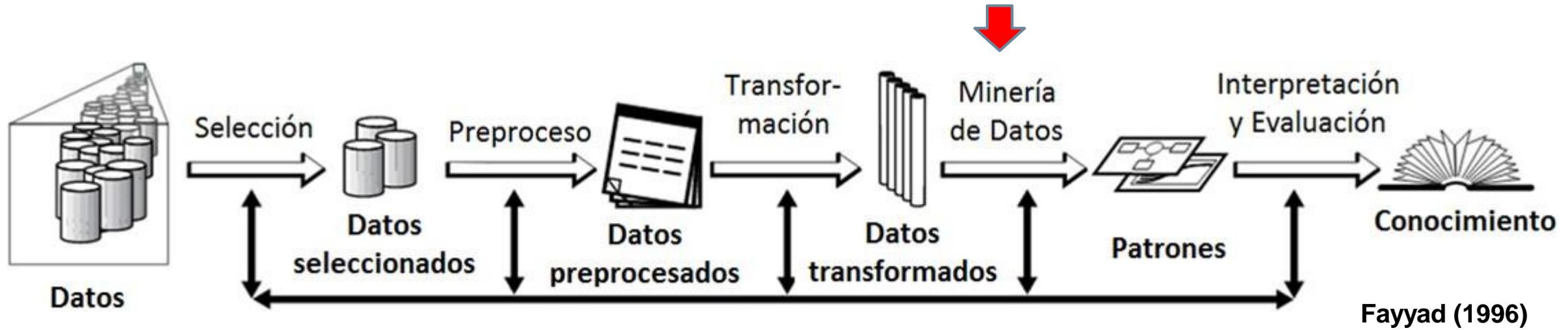
- Elegidos en base al problema
- Medidas subjetivas y objetivas

Extracción de Conocimiento (proceso KDD)

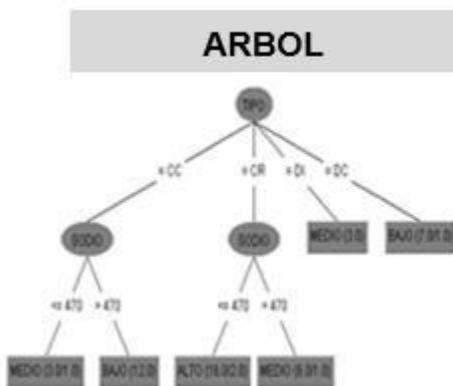


- Uniformar la notación.
- Datos faltantes
- Fuera de los rangos esperados (outliers)

Extracción de Conocimiento (proceso KDD)



□ Técnicas de Minería de Datos



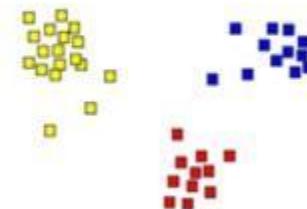
REGLAS

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

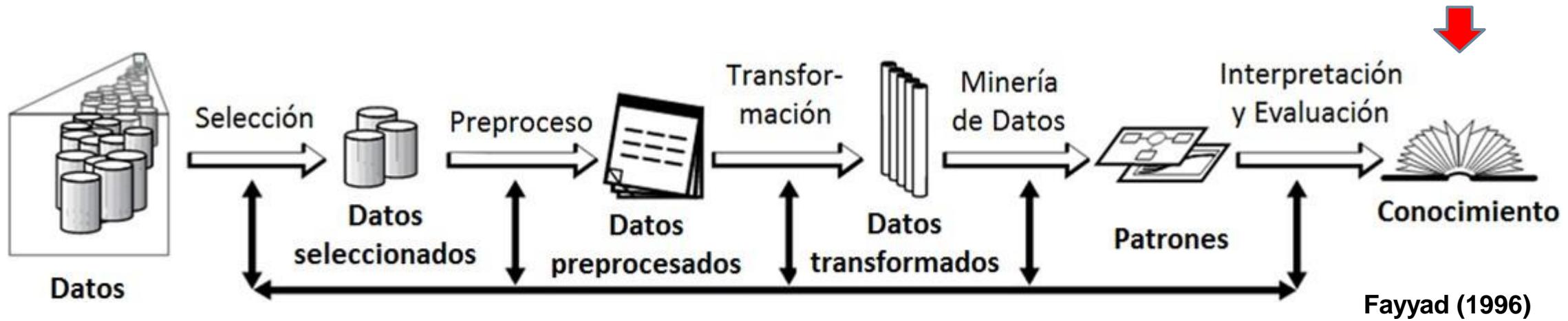
IF (TIPO = CR) AND (PRODUCTO = CN)
ENTONCES (COSTO=ALTO)

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

AGRUPAMIENTO



Extracción de Conocimiento (proceso KDD)



Análisis inteligente

Técnicas de representación



Minería de Datos vs otras disciplinas

- Los sistemas tradicionales de explotación de datos están basados en la existencia de hipótesis o modelos previos.
- Problemas
 - Quien formula la hipótesis debe saber cuál es la información que necesita.
 - La complejidad de los datos almacenados y sus interrelaciones dificulta la verificación del modelo.
- La Minería de Datos busca el descubrimiento del conocimiento **sin una hipótesis** preconcebida.

Ej. 1: Resultado de un curso

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

Ej. 1: Resultado de un curso

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				
0		SI (I NGRESO = APROB)		entonces (RESULT=APROB)
12				
65		SI (I NGRESO = DESAP) AN D		
75		(F ORO = NO)	entonces	(RESULT=DESAP)
60				
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

Tipo de conocimiento a extraer

13

Predictivo

- Enbase al modelo construido es posible predecir hechos futuros.
- Por ejemplo se busca predecir:
 - Cuál medicamento suministrar a un paciente dado.
 - Si un mail recibido es spam o no.

Descriptivo

- Muestran nuevas relaciones entre las variables.
- Por ejemplo se buscará describir:
 - Tipos de clientes para diseñar campañas de marketing
 - Transacciones en una tarjeta de crédito para detectar casos anómalos.

Tarea predictiva - Aprendizaje supervisado

GATO



GATO



GATO



ARBOL



ARBOL



CUADERNO



CUADERNO



CUADERNO



GATO



?

Ejemplo de tarea predictiva: Prescripción de lentes

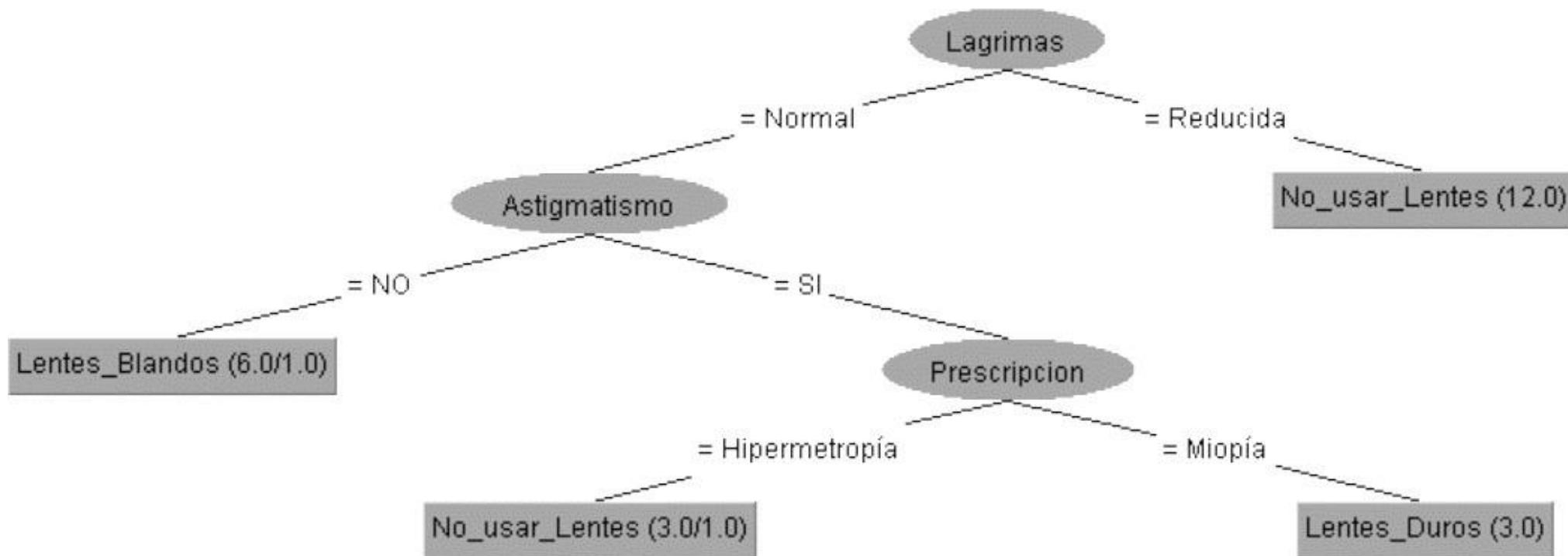
- Se dispone de la siguiente información de pacientes atendidos previamente.
 - **EDAD** del paciente: joven, pre-presbicia, presbicia
 - **PRESRICPCION** de lentes: miope, hipermetrópe
 - **ASTIGMATISMO**: si, no
 - Tasa de producción de **LAGRIMAS**: reducida, normal.
 - **DIAGNOSTICO**
 - el paciente debe usar lentes de contacto duras
 - el paciente debe usar lentes de contacto blandas
 - el paciente no debe usar lentes de contacto.

Conjunto de ejemplos etiquetados

Id	Edad	Espectativa	Astigmatismo	Lagrimas	Diagnóstico
1	Joven	Hipermetropía	NO	Normal	Lentes_Blandos
2	Joven	Miopía	NO	Normal	Lentes_Blandos
3	Joven	Hipermetropía	SI	Normal	Lentes_Duros
4	Joven	Miopía	SI	Normal	Lentes_Duros
5	Joven	Hipermetropía	NO	Reducida	No_usar_Lentes
...
...
22	Presbicia	Miopía	NO	Reducida	No_usar_Lentes
23	Presbicia	Miopía	NO	Normal	No_usar_Lentes
24	Presbicia	Miopía	SI	Reducida	No_usar_Lentes

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

Arbol de Clasificación



Tarea descriptiva - Aprendizaje no supervisado



AGRUPAMIENTO

Ejemplo de tarea descriptiva: Caracterización de flores

- Se dispone de información de 3 tipos de flores Iris



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

Tarea descriptiva: Caracterización de flores

Id	sepallength	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

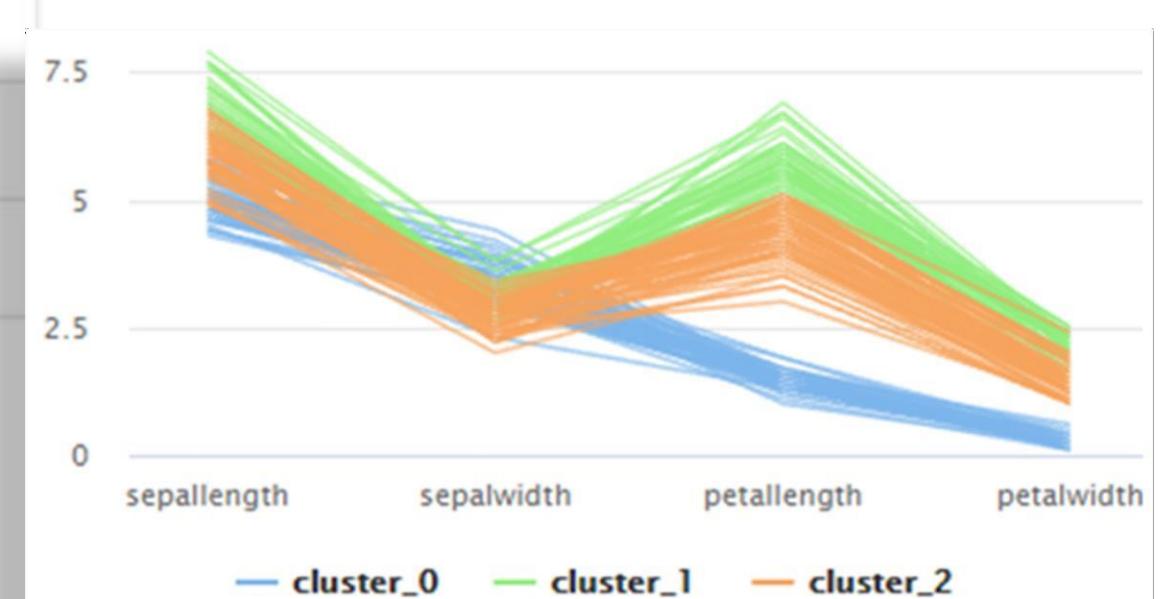
<https://archive.ics.uci.edu/ml/datasets/Iris>

Tarea descriptiva: Caracterización de flores

□ Resultado de agrupar los ejemplos

Attribute	cluster_0	cluster_1	cluster_2
sepallength	5.006	6.854	5.884
sepalwidth	3.418	3.077	2.741
petallength	1.464	5.715	4.389
petalwidth	0.244	2.054	1.434

Identificar valores distintos en los diferentes grupos



Minería de Datos y el proceso de KDD



Comenzaremos analizando los datos disponibles

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

Tipos de variables

□ Cuantitativas o numéricas

- DISCRETAS(cant. de empleados, cant. de alumnos, etc)
- CONTINUAS (sueldo, metros cuadrados, beneficios, etc)

□ Cualitativas o categóricas

- NOMINALES: nombran al objeto al que se refieren sin poder establecer un orden (estado civil, raza, idioma, etc.)
- ORDINALES: se puede establecer un orden entre sus valores (alto, medio, bajo, etc)

Ejercicio

Premios.csv

- Se dispone de la siguiente información de los premios de la Academia otorgados a los mejores actores y actrices desde 1928 hasta 2016.
 - Año en que fue otorgado el premio
 - Datos del actor que lo recibió: Nombre, edad, sexo
 - Datos de la película: Título, género, duración, rating, cantidad de nominaciones que recibió, mes de estreno, sinopsis

Ejercicio

Premios.csv

- ◆ El archivo **Premios.csv** contiene 178 premios otorgados

Year	Age	Actor	Sex	Film	nominations	rating	duration	genre1	genre2	release	synopsis
1928	22	Laura Gainor	F	Sunrise	5	7.8	110	Drama	Romance		A street cleaner sav
1928	44	Emil Jannings	M	The Last Command	2	8	88	Drama	History	April	A former Imperial R
1929	37	Mary Pickford	F	Coquette		7.3	76	Drama	Romance	April	A flirtatious souther
...
...
2015	26	Brie Larson	F	Room	4	8.2	118	Drama	Thriller	January	ROOM tells the extr
2016	28	Emma Stone	F	La La Land	11	8.6	128	Drama	Romance	December	While navigating th
2016	41	Casey Affleck	M	Manchester by the Sea	6	8.8	137	Drama		January	A depressed uncle i

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

RAPIDMINER STUDIO

HERRAMIENTA DE MINERÍA DE DATOS

<https://rapidminer.com/products/studio/>

RapidMiner studio



- Es un entorno para experimentación de análisis de datos que posee implementadas distintas estrategias de Minería de Datos.
- Es de distribución libre.
- Opera a través de la conexión de componentes visuales.

EJEMPLO : PREMIOS.CSV



- Utilizaremos RapidMiner Studio para analizar la información disponible.
- Antes de comenzar, asegúrese de que dispone del archivo

PREMIOS.CSV

- De no ser así, puede descargarlo del siguiente URL

File Edit Process



Operators

Repository

 Import Data

▶  Training Resources

▶  Samples

▶  Community Sam

▶  DB

Start

Recent

Learn

Start a new project



Blank

Start a new process from scratch in the
design view.

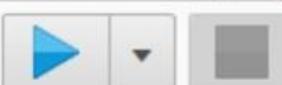


Turbo Pr

Prepare yo
transform,



Comenzaremos con un
proyecto en blanco



Views:

Design

Results

Turbo Prep

More

Find data, operators...etc



All Studio

Operators X

Repository X

+ Import Data ≡

▶ Training Resources (Laura)

▶ Samples

▶ Community Samples (Laura)

▶ DB

▶ ATAQUE_REDES (Laura)

▶ Local Repository (Laura)

▶ MBBS 2018 (Laura)

▶ MD (postgrado) (Laura)

▶ MD - PRACTICAS (Laura)

▶ MD_Educacion (Laura)

▶ MIDUSI (profesor)

▶ MINERIA (Laura)

▶ Mineria2018 (Laura)

▶ Representacion2018 (Laura)

< >

Process Help

● Process 100%

Process

inp res

Your process looks empty.
Add some data first.
Drag data or operators here.

Recommended Operators i

Retrieve 12% Select Attributes 6% Set Role 5%

Parameters X

Process

logverbosity init

logfile

resultfile

random seed 2001

send mail never

encoding SYSTEM

[Hide advanced parameters](#)

✓ [Change compatibility \(9.2.000\)](#)



Views:

Design

Results

Turbo Prep

More ▾

Find data, operators...etc



All Studio ▾

Repository ×

Operators ×

read ×

▼ Data Access (24)

▼ Files (16)

▼ Read (15)

Read CSV ▲

Read Excel

Read Excel with header

Read URL

Read SPSS

Read Stata

Read Sparse

We found "Spreadsheet Table Extraction", "SAS Connector" and one more result in the Marketplace. [Show me!](#)

Process Help ×

● Process 100%

Process

inp res

Utilice doble-click sobre el operador o arrastre y suelte en el área del proceso

Recommended Operators ⓘ

Retrieve 12%

Select Attributes 6%

Set Role 5%

Parameters ×

Process

logverbosity **init** ▾

logfile

resultfile

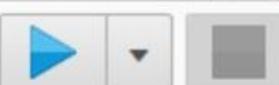
random seed **2001**

send mail **never** ▾

encoding **SYSTEM** ▾

[Hide advanced parameters](#)

[Change compatibility \(9.2.000\)](#)



Views:

Design

Results

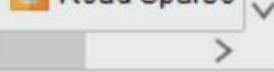
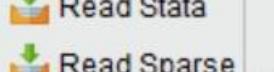
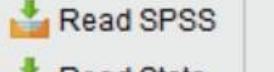
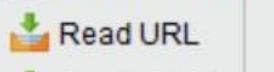
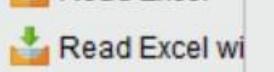
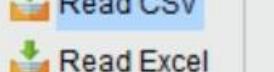
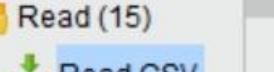
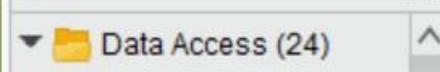
Turbo Prep

More

Find data, operators...etc



All Studio



We found "Spreadsheet Table Extraction", "SAS Connector" and one more result in the Marketplace. [Show me!](#)

Click to select, drag to move.

Process

Help

Process

100%



Read CSV



Conecte el operador

Recommended Operators



Select Attributes

35%

Set Role

32%

Apply Model

25%

Parameters

Read CSV

Import Configuration Wizard...

csv file



column separator:

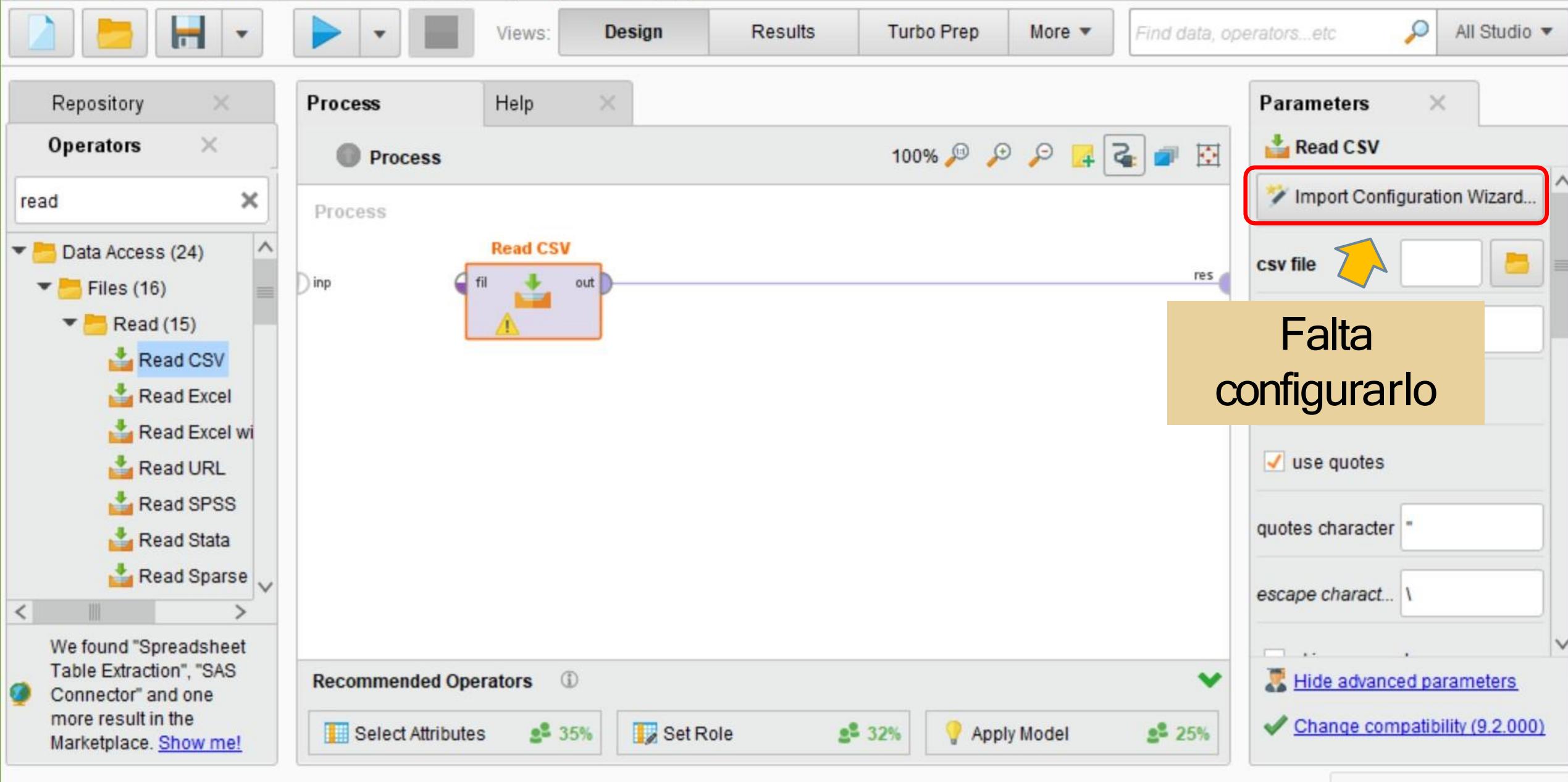
 trim lines use quotes

quotes character:



escape character:

[Hide advanced parameters](#)[Change compatibility \(9.2.000\)](#)



Select the data location.

DatosCSV

Bookmarks	File Name	Size	Type	Last Modified
★ --- Last Directory	Ataques a Redes	8 KB	File Folder	Feb 27, 2019
	Drug5.csv	5 KB	Archivo de valores s...	Sep 27, 2018
	Drug5_numerico.csv	4 KB	Archivo de valores s...	Sep 27, 2018
	iris.csv	1 KB	Archivo de valores s...	Aug 9, 2017
	lentes.csv	36 KB	Archivo de valores s...	Feb 25, 2019
	Premios.csv	85 KB	Archivo de valores s...	Mar 1, 2019
	Sonar.csv			Feb 27, 2019

PREMIOS.CSV

Premios.csv

CSV (.tsv, .csv)

← Previous → Next X Cancel

Select the data location.

Bookmarks		File Name	Size	Type	Last Modified
★	--- Last Directory	Ataques a Redes		File Folder	Feb 27, 2019
		Drug5.csv	8 KB	Archivo de valores s...	Sep 27, 2018
		Drug5_numerico.csv	5 KB	Archivo de valores s...	Sep 27, 2018
		iris.csv	4 KB	Archivo de valores s...	Aug 9, 2017
		lentes.csv	1 KB	Archivo de valores s...	Feb 25, 2019
		Premios.csv	36 KB	Archivo de valores s...	Mar 1, 2019
		Sonar.csv	85 KB	Archivo de valores s...	Feb 27, 2019

Premios.csv

CSV (.tsv, .csv)



← Previous

→ Next

Cancel

Specify your data format

 Header Row

1

File Encoding

windows-1252

 Use Quotes

"

Start Row

1

Escape Character

\

 Trim Lines

Column Separator

Comma ";"

Decimal Character

.

 Skip Comments

#

1	Year	Age	Actor	Sex	Film	nominati...	rating	duration	genre1	genre2
2	1928	44	Emil Jan...	M	The Last...	2	8	88	Drama	History
3	1928	22	Laura G...	F	Sunrise	5	7.8	110	Drama	Romanc
4	1929	38	Warner ...	M	In Old Ari...	5	5.8	95	Romance	Western
5	1929	37	Mary Pic...	F	Coquette		7.3	76	Drama	Romanc
6	1930	62	George ...	M	Disraeli	3	6.5	90	Biography	Drama
7	1930	30	Norma S...	F	The Divo...	4	6.9	84	Romance	Drama
8	1931	53	Lionel B...	M	A Free S...	3	6.7	93	Crime	Drama
9	1931	62	Marie Dr...	F	Min and ...		7.4	69	Comedy	Drama
10	1932	41	W. Beery...	M	The Cha...	4	7.3	86	Drama	Family
11	1932	32	Helen H...	F	Sin of Ma...		6.8	75	Drama	

 no problems.[← Previous](#)[Next →](#)[Cancel X](#)

Format your columns.

Date format

 Replace errors with missing values (i)

	Year <i>integer</i>	Age <i>integer</i>	Actor <i>polynominal</i>	Sex <i>polynominal</i>	Film <i>polynominal</i>	nominations <i>integer</i>	Run time <i>integer</i>
1	1928	44	Emil Jannings	M	The Last Comm...	2	85
2	1928	22	Laura Gainor (ak...	F	Sunrise	5	75
3	1929	38	Warner Baxter	M	In Old Arizona	5	85
4	1929	37	Mary Pickford	F	Coquette	?	75
5	1930	62	George Arliss	M	Disraeli	3	65
6	1930	30	Norma Shearer	F	The Divorcee	4	65
7	1931	53	Lionel Barrymore	M	A Free Soul	3	65
8	1931	62	Marie Dressler	F	Min and Bill	?	75
9	1932	41	W. Beery(47)/F. M...	M	The Champ/Dr. J...	4	75
10	1932	32	Helen Hayes	F	Sin of Madelon	?	65
11	1933	34	Charles Laughton	M	Private Life Henry...	2	75
12	1933	26	John Barrymore	F	Missing Links	0	75



no problems.

Previous

Finish

Cancel

<new process*> – RapidMiner Studio Free 9.2.000 @ LauraUB

File Edit Process View Connections Cloud Settings Extensions Help

Views: Design Results Turbo Prep More Find data, operators...etc All Studio

Repository Operators

read

Data Access (24)

Files (16)

Read (15)

Read CSV

Read Excel

Read Excel wi

Read URL

Read SPSS

Read Stata

Read Sparse

We found "Spreadsheet Table Extraction", "SAS Connector" and one more result in the Marketplace. [Show me!](#)

Process Help

Process

100%

Process

inp → Read CSV → res

res

Ejecutar

Parameters

Read CSV

Import Configuration Wizard...

csv file nios.csv

column separator ,

trim lines

use quotes

quotes character "

escape character \

[Hide advanced parameters](#)

[Change compatibility \(9.2.000\)](#)

Recommended Operators

Select Attributes 35% Set Role 32% Apply Model 25%

```
graph LR; inp((inp)) --> ReadCSV[Read CSV]; ReadCSV --> res1((res)); res1 --> res2((res))
```

<new process*> – RapidMiner Studio Free 9.2.000 @ LauraUB

File Edit Process View Connections Cloud Settings Extensions Help

Views: Design Results Turbo Prep More Find data, operators...etc All Studio

Result History ExampleSet (Read CSV) X

Open in Turbo Prep

Roles: all

Utilícelos para cambiar entre la vista de diseño y la de resultados

Row No.	Year				nominations	
1	1928				2	
2	1928	22	Laura Gainor ...	F	Sunrise	5
3	1929	38	Warner Baxter	M	In Old Arizona	5
4	1929	37	Mary Pickford	F	Coquette	?
5	1930	62	George Arliss	M	Disraeli	3
6	1930	30	Norma Shear...	F	The Divorcee	4
7	1931	53	Lionel Barry...	M	A Free Soul	3
8	1931	62	Marie Dressler	F	Min and Bill	?
9	1932	41	W. Beery(47)/...	M	The Champ/...	4

ExampleSet (178 examples, 0 special attributes, 12 regular attributes)

Repository

- + Import Data
- Training Resources (conn)
- Samples
- Community Samples (cor)
- DB
- ATAQUE_REDES (Laura)
- Local Repository (Laura)
- MBBS 2018 (Laura)
- MD (postgrado) (Laura)
- MD - PRACTICAS (Laura)
- MD_Educacion (Laura)
- MIDUSI (profesor)
- MINERIA (Laura)
- Mineria2018 (Laura)
- Representacion2018 (Lau
- Tesina_AR (Laura)

<new process*> – RapidMiner Studio Free 9.2.000 @ LauraUB

File Edit Process View Connections Cloud Settings Extensions Help

Views: Design Results Turbo Prep More Find data, operators...etc All Studio

Result History ExampleSet (Read CSV) Click for more options

Metadatos

	Type	Missing	St...	Filter (12 / 12 attributes):	Search for Attributes	▼
Year	Integer	0	Min 1928	Max 2016	Avg 19	▼
Age	Integer	0	Min 21	Max 81	Avg 39	▼
Actor	Polynomial	0	Least Yul Brenner (1)	Most Daniel Day-Lewis (3)	Var	Da
Sex	Polynomial	0	Least M (89)	Most F (89)	Var	F
Film	Polynomial	0	Least Yankee Doodle Dandy (1)	Most As Good As It Gets (2)	Var	As

Showing attributes 1 - 12 Examples: 178 Special Attributes: 0 Regular Attributes: 12

Repository

- Import Data
- Training Resources (conn)
- Samples
- Community Samples (cor)
- DB
- ATAQUE_REDES (Laura)
- Local Repository (Laura)
- MBBS 2018 (Laura)
- MD (postgrado) (Laura)
- MD - PRACTICAS (Laura)
- MD_Educacion (Laura)
- MIDUSI (profesor)
- MINERIA (Laura)
- Mineria2018 (Laura)
- Representacion2018 (Lau
- Tesina_AR (Laura)

<new process*> – RapidMiner Studio Free 9.2.000 @ LauraUB

File Edit Process View Connections Cloud Settings Extensions Help

Views: Design Results Turbo Prep More Find data, operators...etc All Studio

Result History ExampleSet (Read CSV) X

Filter (12 / 12 attributes): Search for Attributes

Repository Import Data

Training Resources (conn) Samples Community Samples (cor) DB ATAQUE_REDES (Laura) Local Repository (Laura) MBBS 2018 (Laura) MD (postgrado) (Laura) MD - PRACTICAS (Laura) MD_Educacion (Laura) MIDUSI (profesor) MINERIA (Laura) Mineria2018 (Laura) Representacion2018 (Lau) Tesina_AR (Laura)

Al clickear sobre el atributo muestra más información

Statistics

Visualizations

Annotations

Age Integer 0

Actor Polynomial 0

Sex Polynomial 0

Least Yul Brenner (1)

Most Daniel Day-Lewis (3)

M (89)

F (89)

Open visualizations

Acceso al gráfico

Showing attributes 1 - 12 Examples: 178 Special Attributes: 0 Regular Attributes: 12

<new process*> – RapidMiner Studio Free 9.2.000 @ LauraUB

File Edit Process View Connections Cloud Settings Extensions Help

Views: Design Results Turbo Prep More Find data, operators...etc All Studio

Result History

Data

Statistics

Visualizations

Annotations

General

Si se trata de una variable cuantitativa muestra el histograma

Plot 1

Plot type: Histogram

Value columns: Age

Color: -

Number of Bins: 10

Plot style ➤

Add new plot

Frequency

Age

Repository

Import Data

Training Resources (conn)

Samples

Community Samples (conn)

DB

ATAQUE_REDES (Laura)

Local Repository (Laura)

MBBS 2018 (Laura)

MD (postgrado) (Laura)

MD - PRACTICAS (Laura)

MD_Educacion (Laura)

MIDUSI (profesor)

MINERIA (Laura)

Mineria2018 (Laura)

Representacion2018 (Laura)

Tesina_AR (Laura)

Age Bin	Frequency
[25, 30]	~15
[30, 35]	~32
[35, 40]	~45
[40, 45]	~37
[45, 50]	~22
[50, 55]	~13
[55, 60]	~13
[60, 65]	~11
[70, 75]	~1
[75, 80]	~2

<new process*> – RapidMiner Studio Free 9.2.000 @ LauraUB

File Edit Process View Connections Cloud Settings Extensions Help

Views: Design Results Turbo Prep More Find data, operators...etc All Studio

Result History ExampleSet (Read CSV) X

Name Type Missing St... Filter (12 / 12 attributes): Search for Attributes

rating Real 0 5.800 9.200

duration Integer 0 Min 69 Max 238 Avg 111

genre1 Polynominal 0 Least Thriller (1)

genre2 Polynominal 35 Least War (1) Most Drama (74)

release Polynominal

Showing attributes 1 - 12 Examples: 178 Special Attributes: 0 Regular Attributes: 12

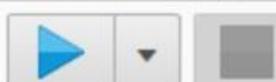
Repository Import Data

Training Resources (conn) Samples Community Samples (cor) DB ATAQUE_REDES (Laura) Local Repository (Laura) MBBS 2018 (Laura) MD (postgrado) (Laura) MD - PRACTICAS (Laura) MD_Educacion (Laura) MIDUSI (profesor) MINERIA (Laura) Mineria2018 (Laura) Representacion2018 (Lau) Tesina_AR (Laura)

Statistics

Si se trata de una variable cualitativa muestra el diagrama de barras

A yellow arrow points upwards from the text "Si se trata de una variable cualitativa muestra el diagrama de barras" towards the genre1 section of the statistics view.



Views:

Design

Results

Turbo Prep

More ▾

Find data, operators...etc



All Studio ▾

Result History

ExampleSet (Read CSV)



Data



Statistics



Visualizations



Annotations

Plot

Plot 1

Plot type

Bar (Column)

Value columns

genre1

Aggregate data

Group by

genre1

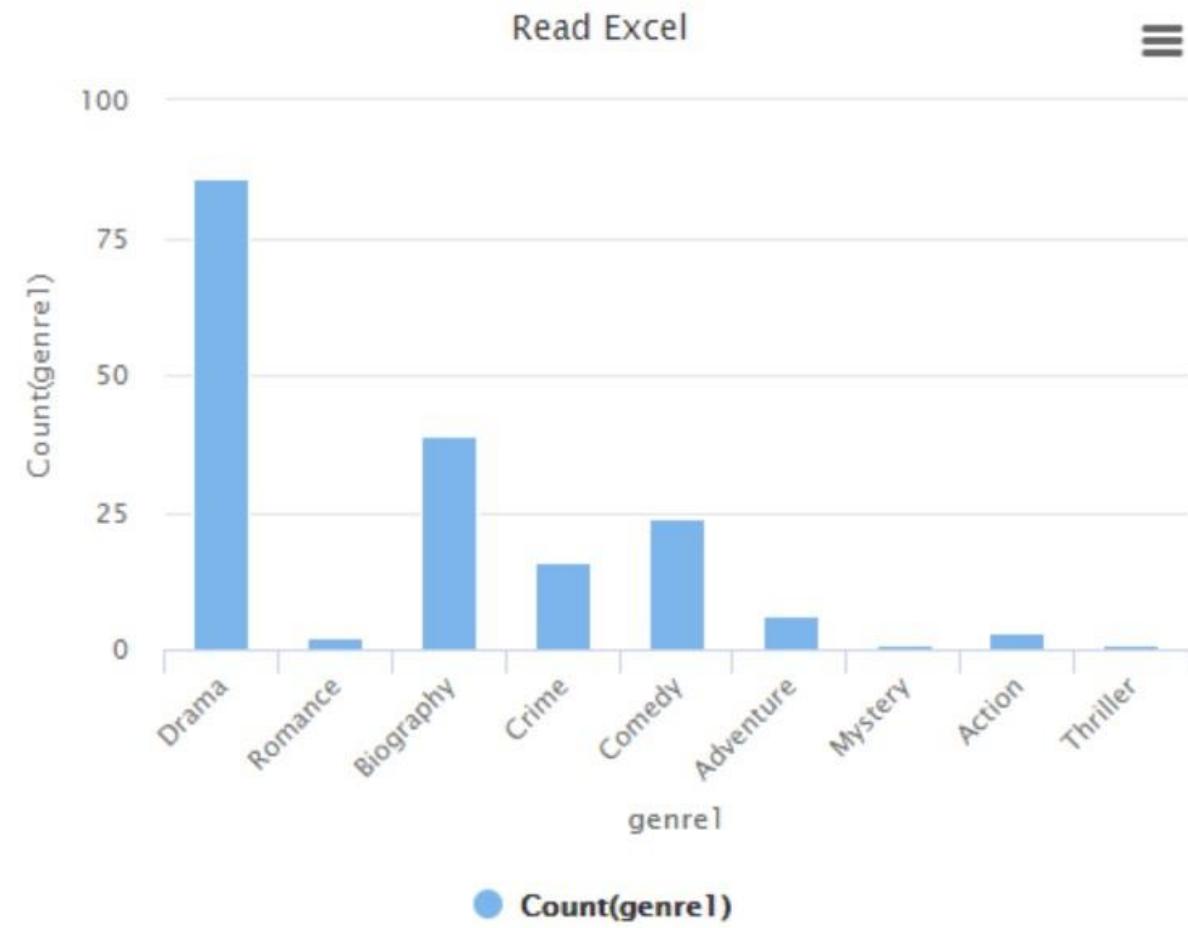
Aggregation Function

Count

Stacking

No stacking

Plot style ➤



Repository

+ Import Data

- ▶ Training Resources (conn)
- ▶ Samples
- ▶ Community Samples (conn)
- ▶ DB
- ▶ ATAQUE_REDES (Laura)
- ▶ Local Repository (Laura)
- ▶ MBBS 2018 (Laura)
- ▶ MD (postgrado) (Laura)
- ▶ MD - PRACTICAS (Laura)
- ▶ MD_Educacion (Laura)
- ▶ MIDUSI (profesor)
- ▶ MINERIA (Laura)
- ▶ Mineria2018 (Laura)
- ▶ Representacion2018 (Laura)
- ▶ Tesina_AR (Laura)

Descripciones estadísticas básicas

- 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
- Varianza
- Desviación estándar
- Rango
- Cuartiles
- Rango Intercuartil

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

$$\bar{x} = \frac{30 + 36 + 47 + 50 + 52 + 52 + 56 + 60 + 63 + 70 + 70 + 110}{12} = \frac{696}{12} = 58$$

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



$$\bar{x} = 58$$

MEDIA TRUNCADA
¿cómo se calcula?
¿para qué sirve?

MEDIANA

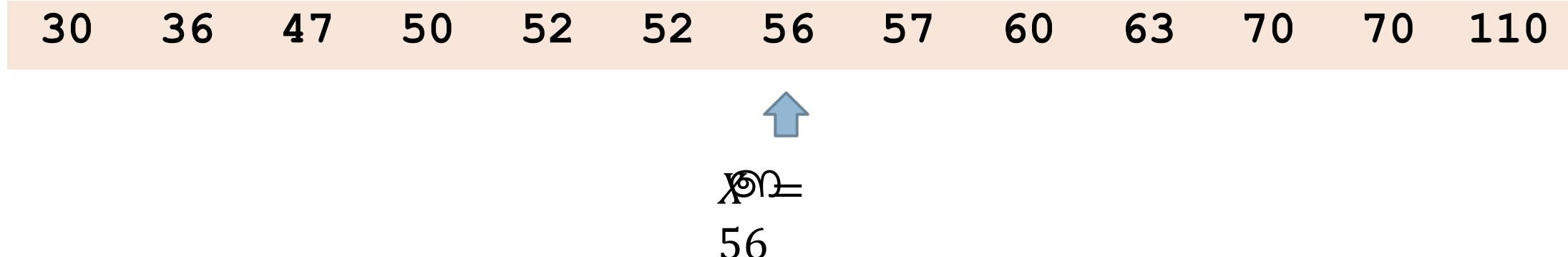
- Divide a los valores del atributo en dos partes iguales de manera que los anteriores son todos menores que él y los siguientes son mayores.
- Antes de calcularla deben **ordenarse los valores** del atributo.
- Ejemplo: atributo numérico con una **cantidad impar** de valores



$$\text{Mediana} = x_{(N+1)/2} = 56$$

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 impar** de valores



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

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

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



$\bar{x}_{50} =$

54

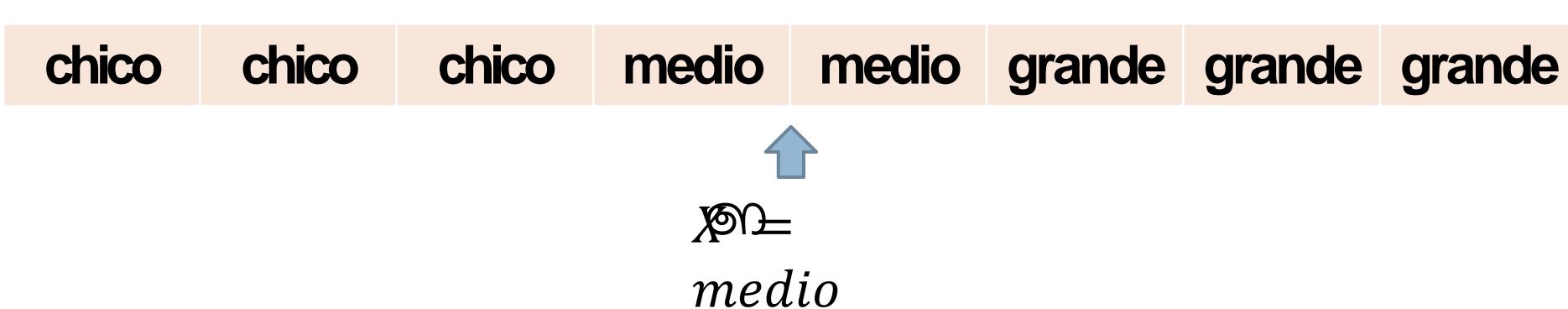
MEDIANA

- 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 impar** de valores



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



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
-------	-------	-------	-------	-------	--------	--------	--------



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.
- Es posible que la mayor frecuencia corresponda a varios valores diferentes, lo que da lugar a más de una MODA.
- Los conjuntos de datos con uno, dos o tres modas se denominan unimodal, bimodal y trimodal, respectivamente.
- 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.

MODA

- La moda es el valor que aparece con mayor frecuencia. Por lo tanto, puede determinarse para atributos cualitativos y cuantitativos.
- 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

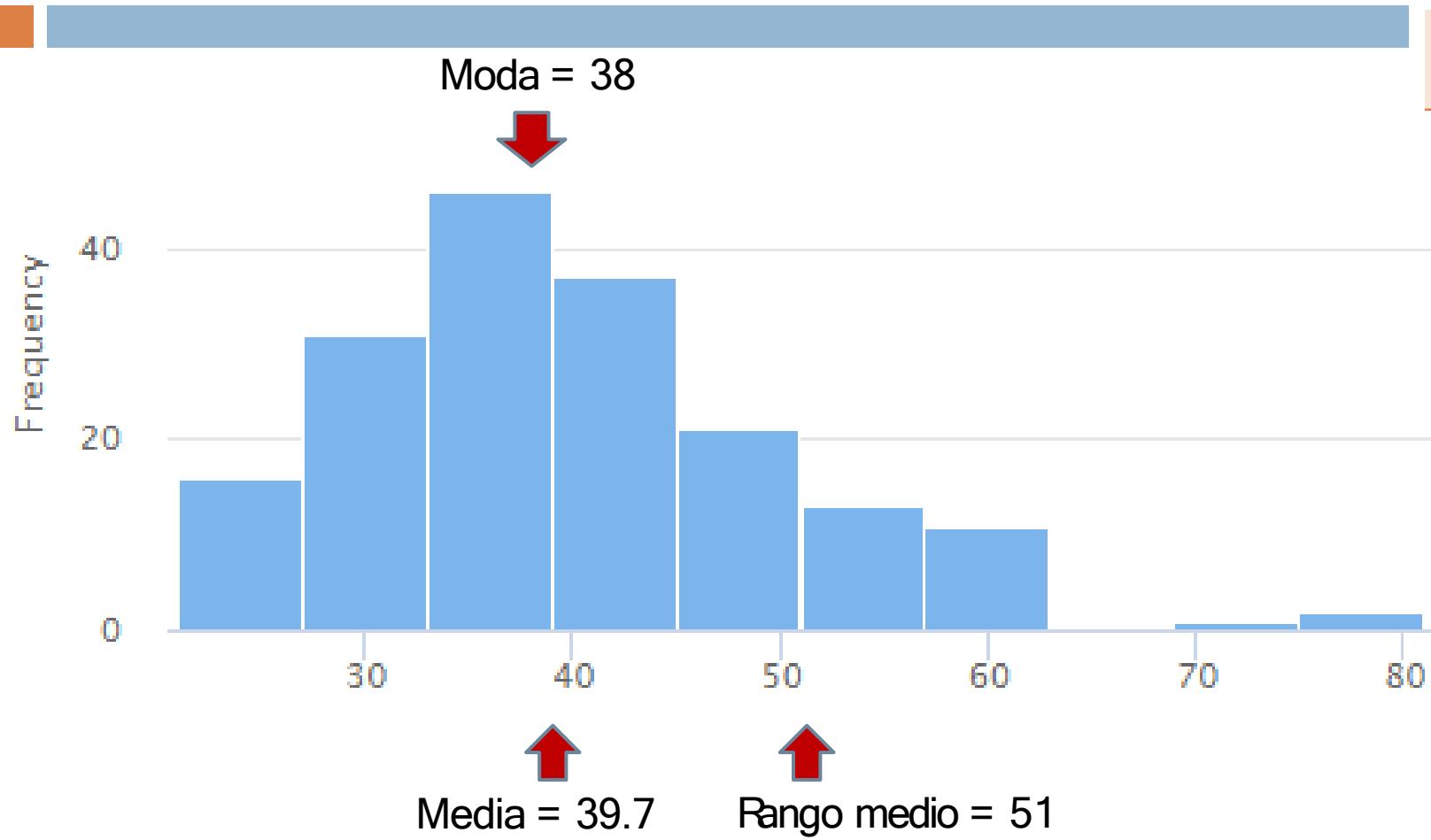
RANGO MEDIO

- 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 AGE

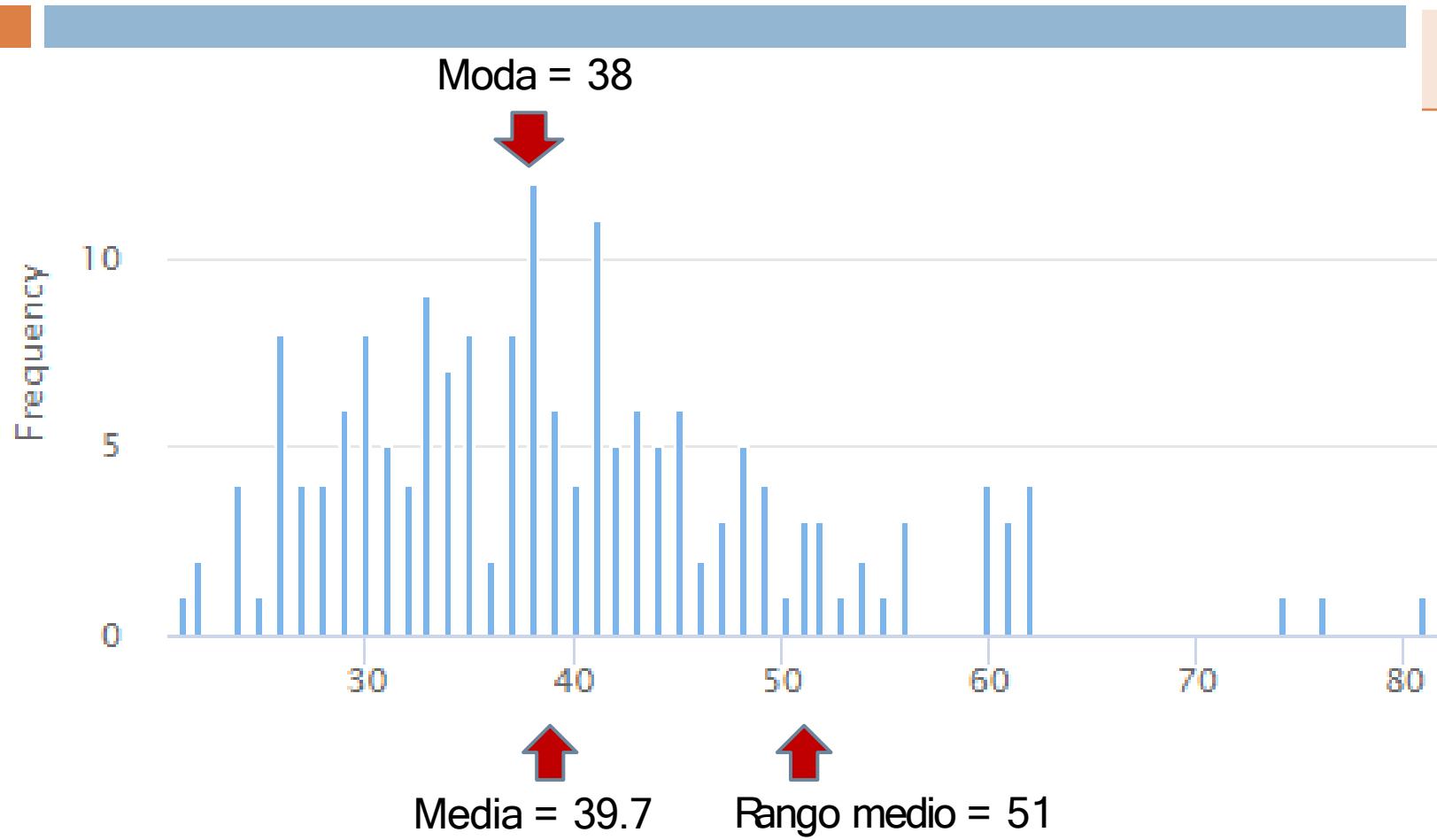


Media	39.702
Mediana	38
Moda	38
Rango medio	51

ID	AGE
1	21
2	22
...	
88	38
89	38
90	38
91	38
...	
177	76
178	81

Mediana = 38

Atributo AGE



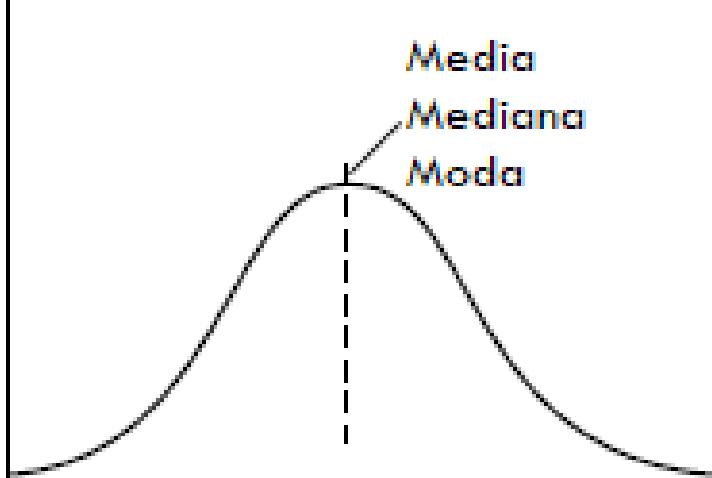
Media	39.702
Mediana	38
Moda	38
Rango medio	51

ID	AGE
1	21
2	22
...	
88	38
89	38
90	38
91	38
...	
177	76
178	81

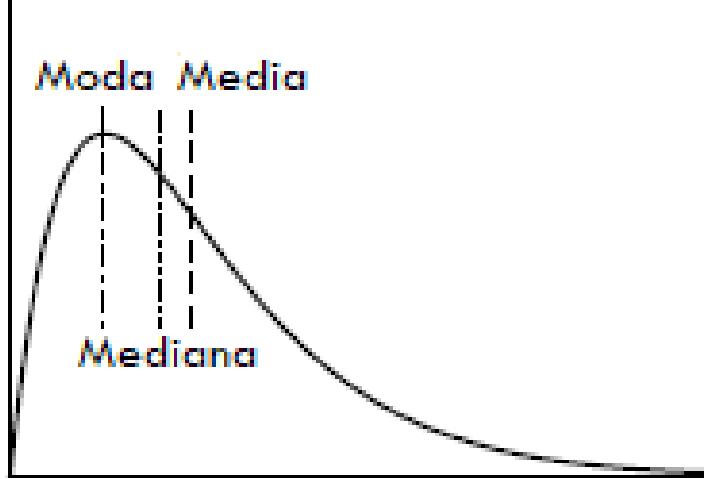
Mediana = 38

Medidas de tendencia central

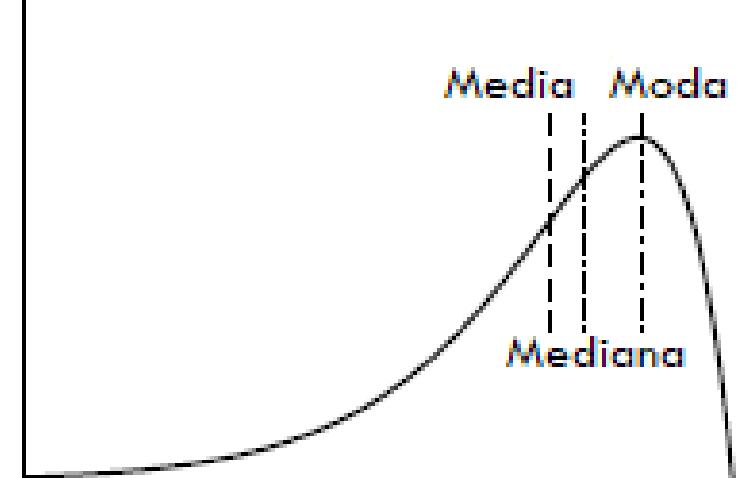
SIMETRICO



SESGO POSITIVO

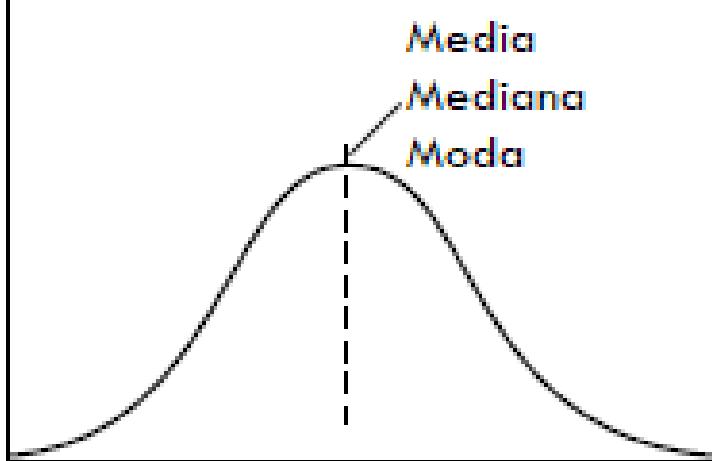


SESGO NEGATIVO

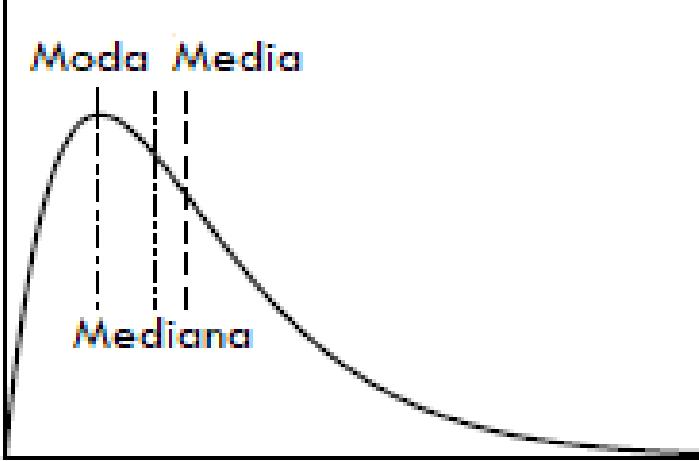


Medidas de tendencia central

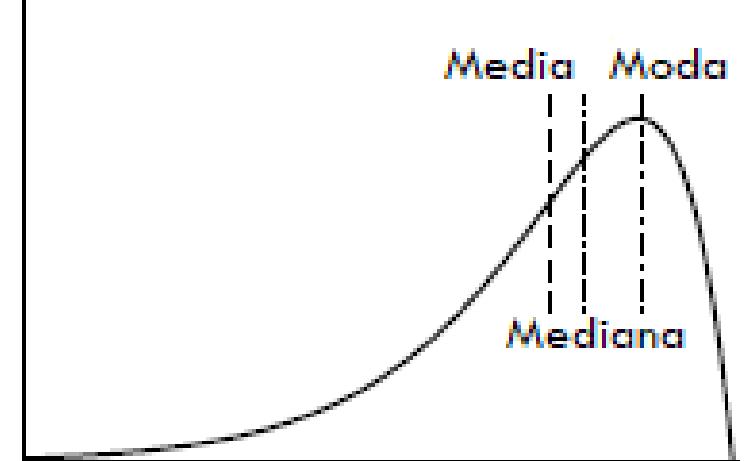
SIMETRICO



SESGO POSITIVO

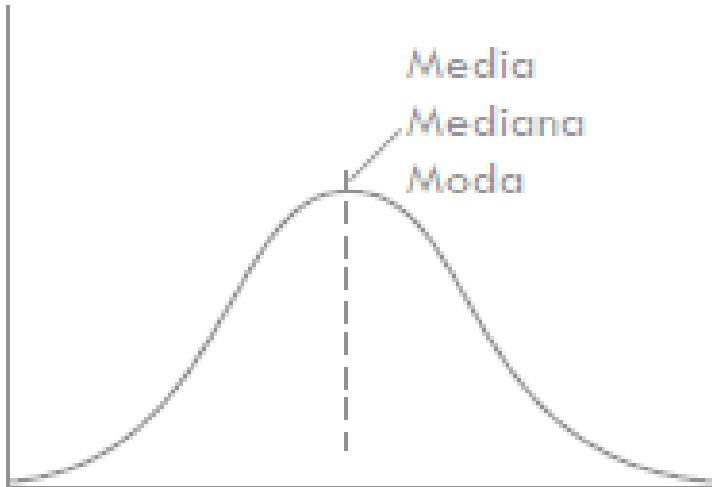


SESGO NEGATIVO

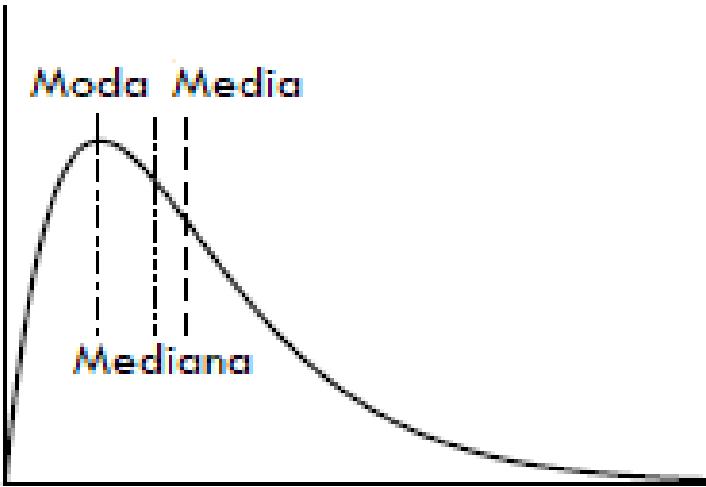


- En una curva de frecuencia unimodal con una distribución de datos perfectamente simétrica, la media, la mediana y la moda coinciden con el valor central.

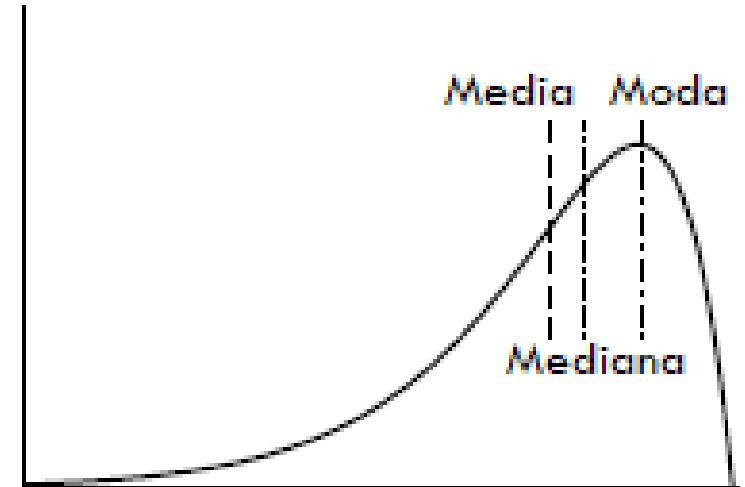
Medidas de tendencia central



SESGO POSITIVO



SESGO NEGATIVO



- El sesgo cuantifica la asimetría de una distribución
- Si $\text{Media} > \text{Moda}$, el sesgo es positivo.
- Si $\text{Media} < \text{Moda}$, el sesgo es negativo.

Descripciones estadísticas básicas

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

- Media
- Mediana
- Moda
- Rango medio

MEDIDAS DE DISPERSION

- Varianza
- Desviación estándar
- Rango
- Cuartiles
- Rango Intercuartil

VARIANZA Y DESVIACION ESTANDARD

- 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**

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

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

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

$E(S^2) = \sigma^2$ es **insesgado**

VARIANZA Y DESVIACION ESTANDARD

□ Ejemplo

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

□ Varianza muestral

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

$$S^2 \approx 413.6364$$

□ Desviación estndar muestral

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

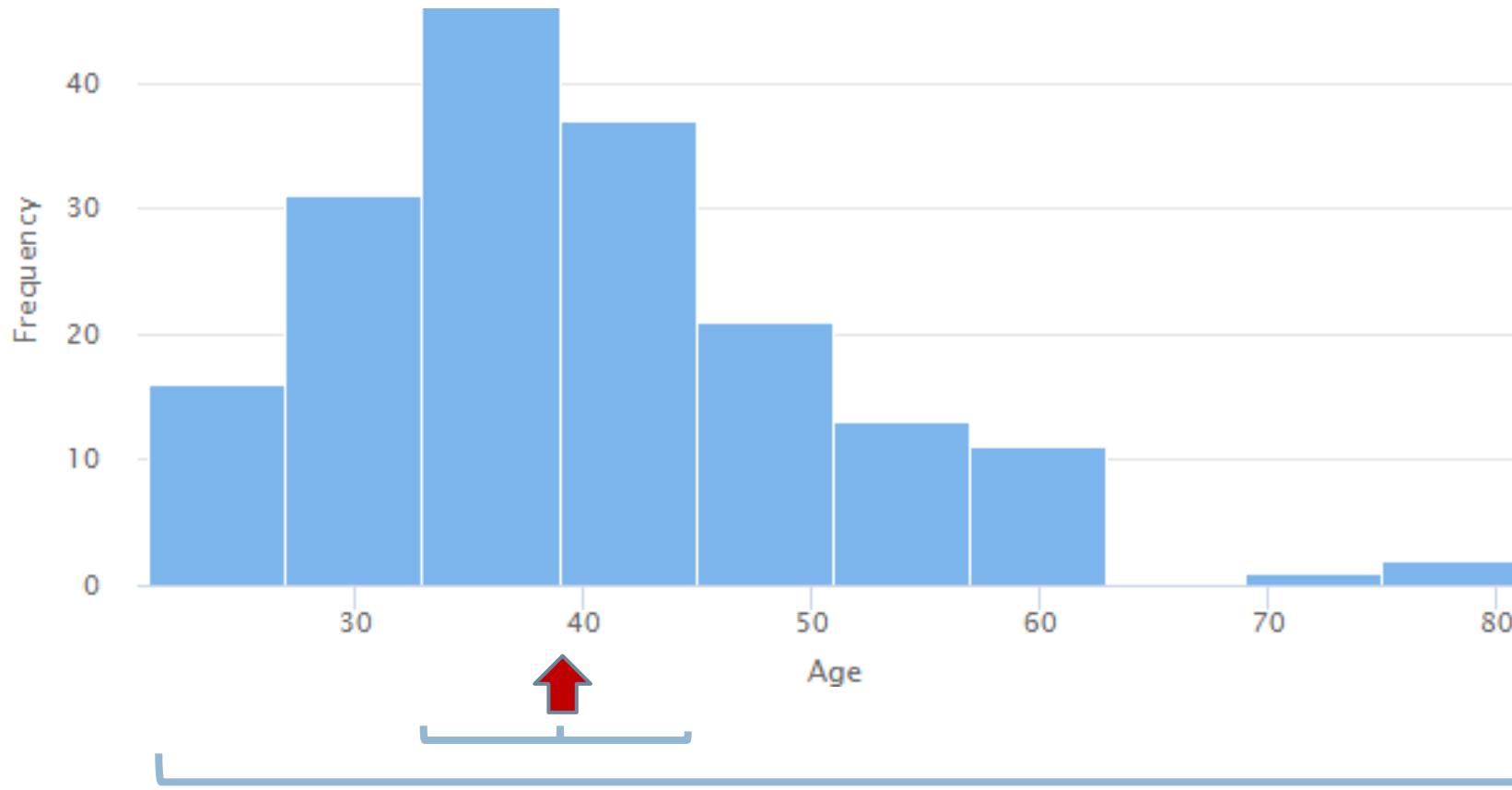
RANGO

- 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 AGE



Media 39.702

Desviación 10.871

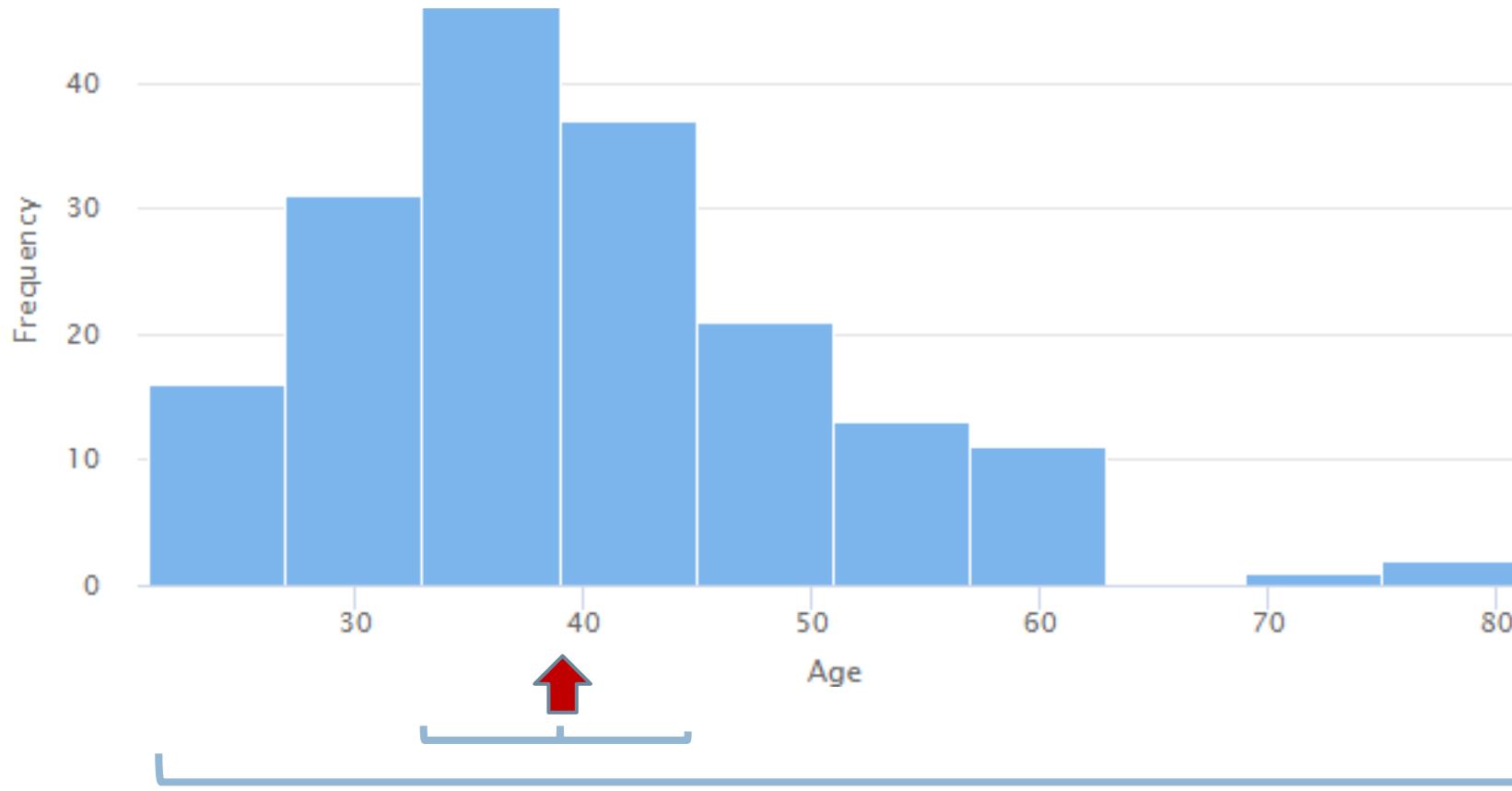
Mínimo 21

Máximo 81

Rango 60

ID	AGE
...	...
175	62
176	74
177	76
178	81

Atributo AGE



Media 39.702

Desviación 10.871

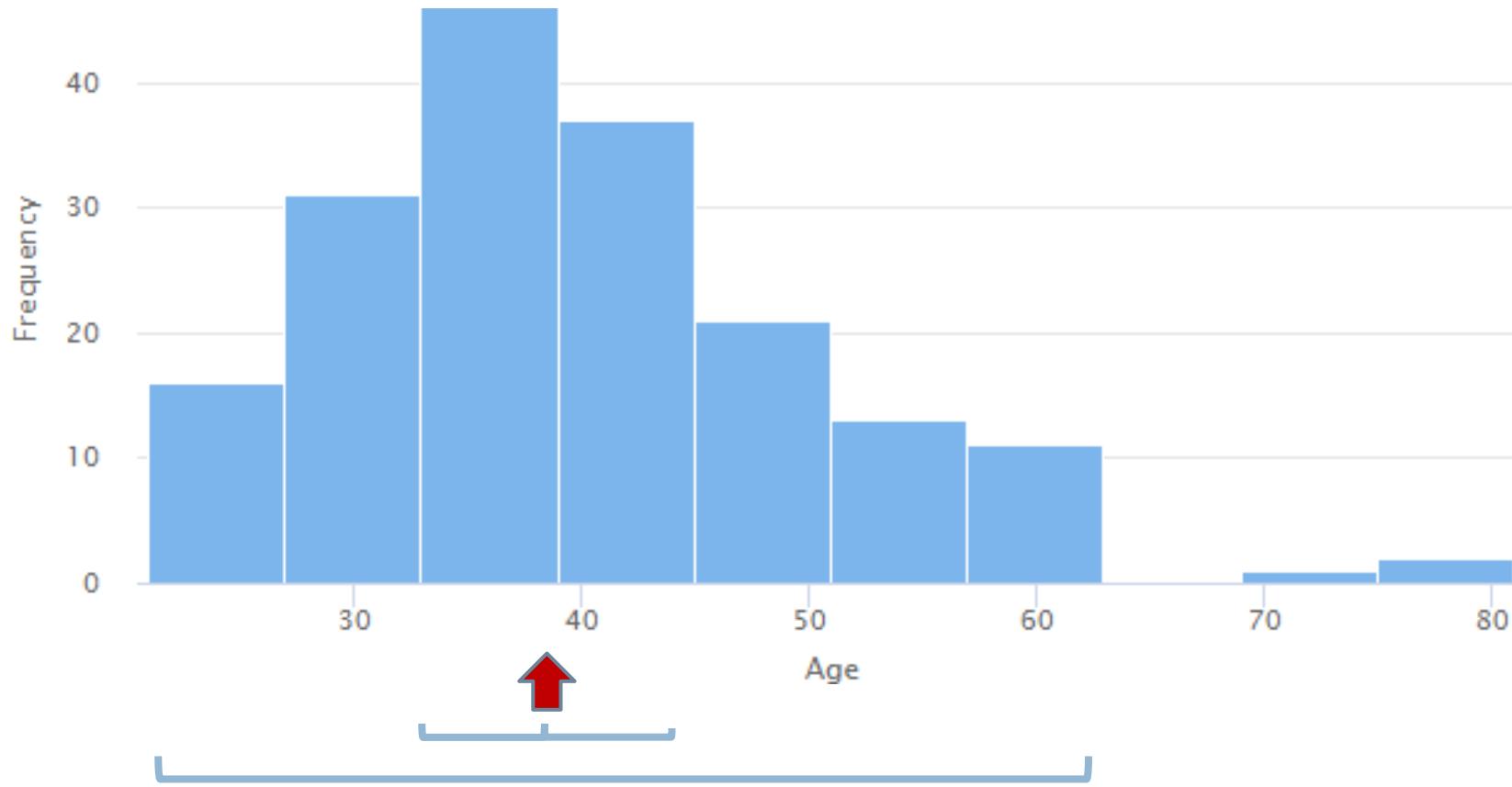
Mínimo 21

Máximo 81

Rango 60

ID	AGE
...	...
175	62
177	76
178	81

Atributo AGE



Media 39.063

Desviación 9.782

Mínimo 21

Máximo 62

Rango 41

ID	AGE
...	...
172	62
173	62
174	62
175	62

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.
- Los cuantiles más usados son los siguientes:
 - 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.*

CUARTILES

- Los cuartiles suelen representarse como Q1, Q2 y Q3. El 2do. cuartil o Q2 coincide con la MEDIANA.
- Usaremos $(N+1)/4$ y $3(N+1)/4$ para hallar las posiciones de Q1 y Q3 respectivamente, siendo N la cantidad de valores disponibles.
 - Si no hay parte decimal, se toma directamente el elemento.
 - Si la posición corresponde a un número con parte decimal entre el elemento i y el $i+1$, se toma el valor proporcional. Sea un número de la forma i,d donde i es la parte entera y d la decimal. El cuartil será:

$$Q = x_i + d (x_{i+1} - x_i)$$

CUARTILES

- Ejemplo:

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

- La ubicación de Q_1 es $(N+1)/4$, es decir, $(12+1)/4=13/4=3.25$
- Como no es un número entero calculamos su valor de manera proporcional entre el 3ro y el 4to valor.

$$\begin{aligned} Q_1 &= x_3 + 0.25 * (x_4 - x_3) \\ &= 47 + 0.25 * (50 - 47) = 47.75 \end{aligned}$$

CUARTILES

- Ejemplo:

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

- La ubicación de Q3 es $3(N+1)/4 = 3*(12+1)/4 = 3*13/4 = 9.75$
- Como no es un número entero calculamos su valor de manera proporcional entre el 9no y el 10mo valor.

$$\begin{aligned}Q_3 &= x_9 + 0.75 * (x_{10} - x_9) \\&= 63 + 0.75 * (70 - 63) = 68.25\end{aligned}$$

CUARTILES

□ Ejemplo:

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



$$Q_1 = 47.75$$



$$Q_2 = 54$$



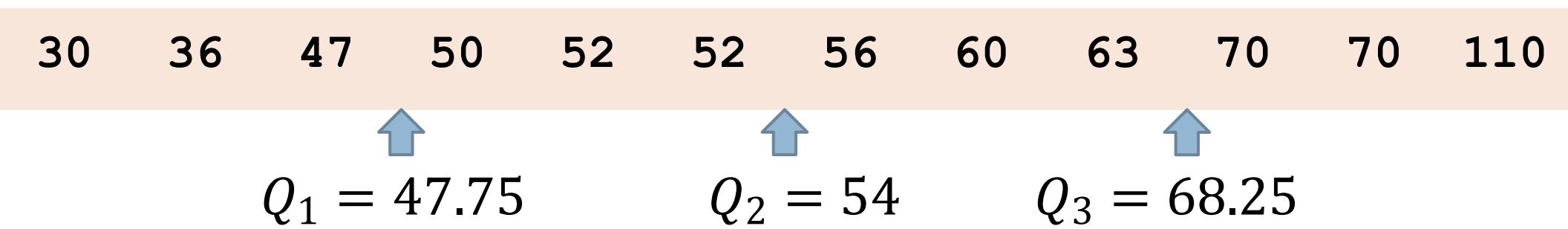
$$Q_3 = 68.25$$

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

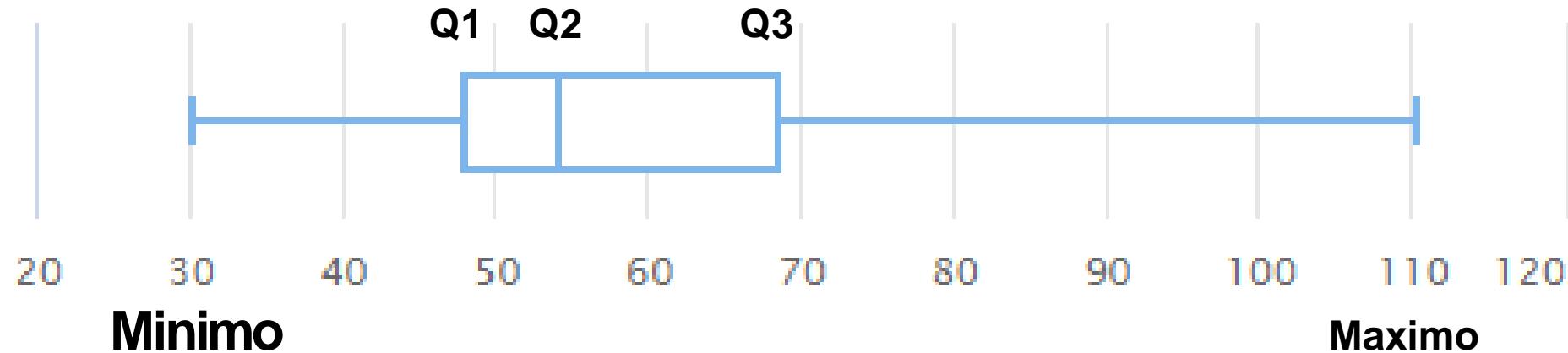
$$\text{RIC} = Q_3 - Q_1$$

- Ejemplo:



$$\text{RIC} = Q_3 - Q_1 = 68.25 - 47.75 = 20.50$$

Diagrama de caja simple



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

Repositorio OpenML

The screenshot shows the OpenML dataset page for 'iris'. At the top, there's a navigation bar with a search bar and a 'HEL' button. Below the header, the dataset name 'iris' is displayed next to a database icon. There are download buttons for ARFF, CSV, and JSON formats. Below these, status information is shown: 'active', 'ARFF', 'Publicly available', 'Visibility: public', and 'Uploaded 06-04-2014'. A large yellow banner at the bottom contains the URL <https://www.openml.org/d/61>.



Id	sepal length	sepal width	petal length	petal width	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

Repositorio OpenML

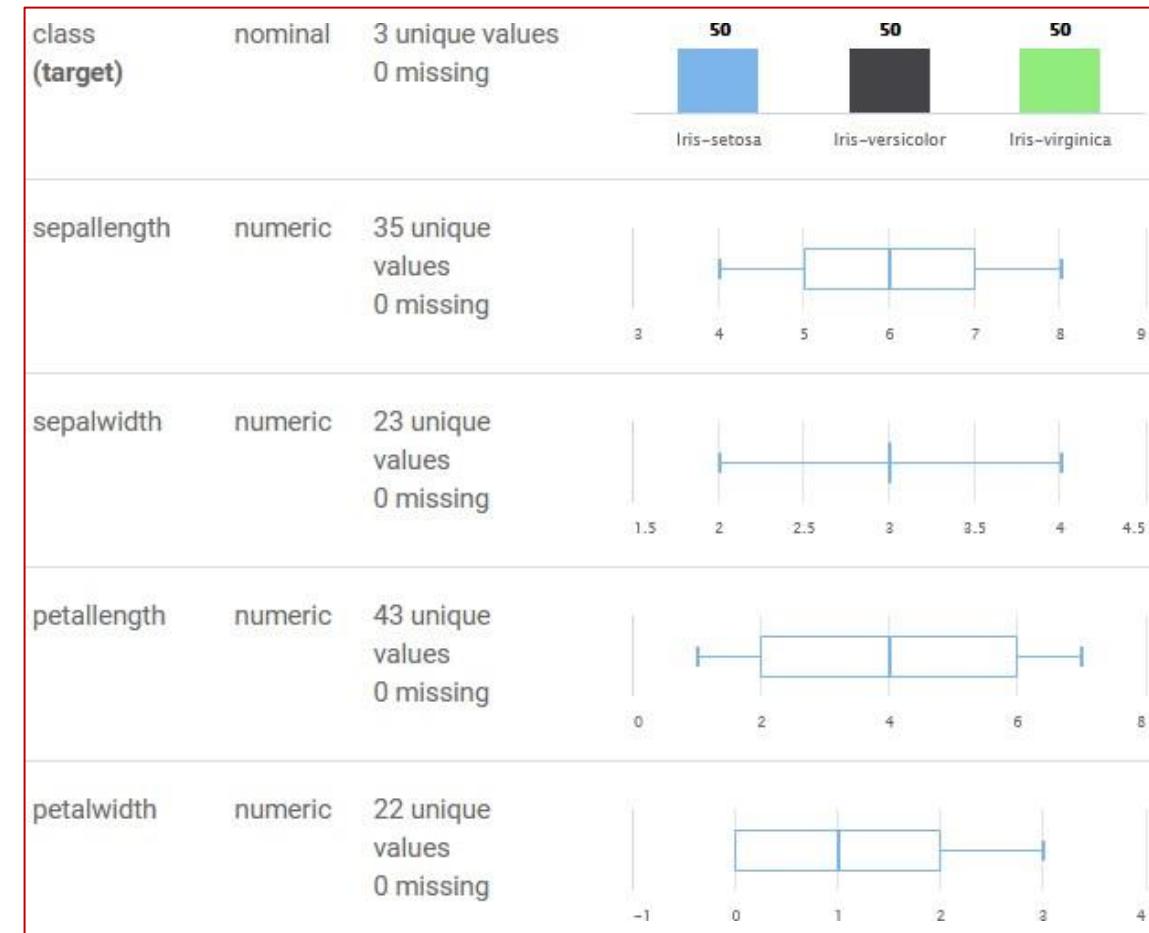
https://www.openml.org/d/61

Search HEL

iris

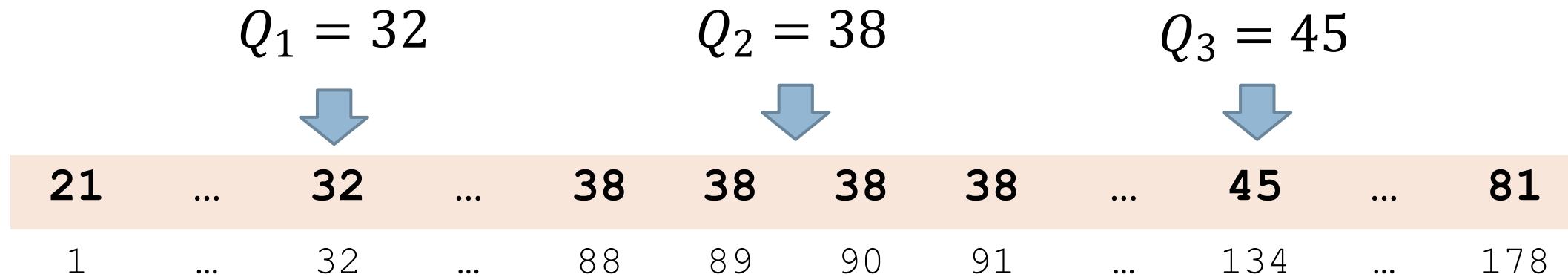
active ARFF Publicly available Visibility: public Uploaded 06-04-2014

<https://www.openml.org/d/61>



Cuartiles y RIC del atributo AGE

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

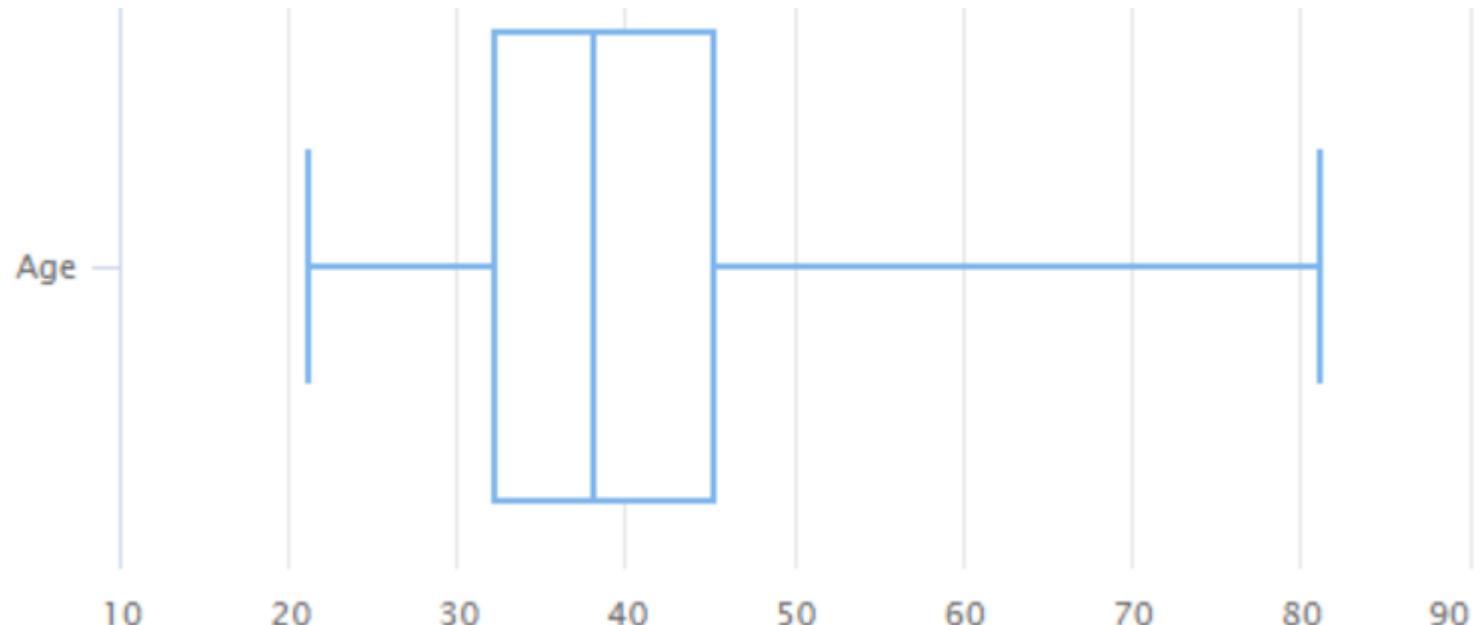


$$RIC = Q_3 - Q_1 = 45 - 32 = 13$$

Diagrama de caja simple

□ Atributo AGE

Minimo	21
Q1	32
Q2	38
Q3	45
Maximo	81



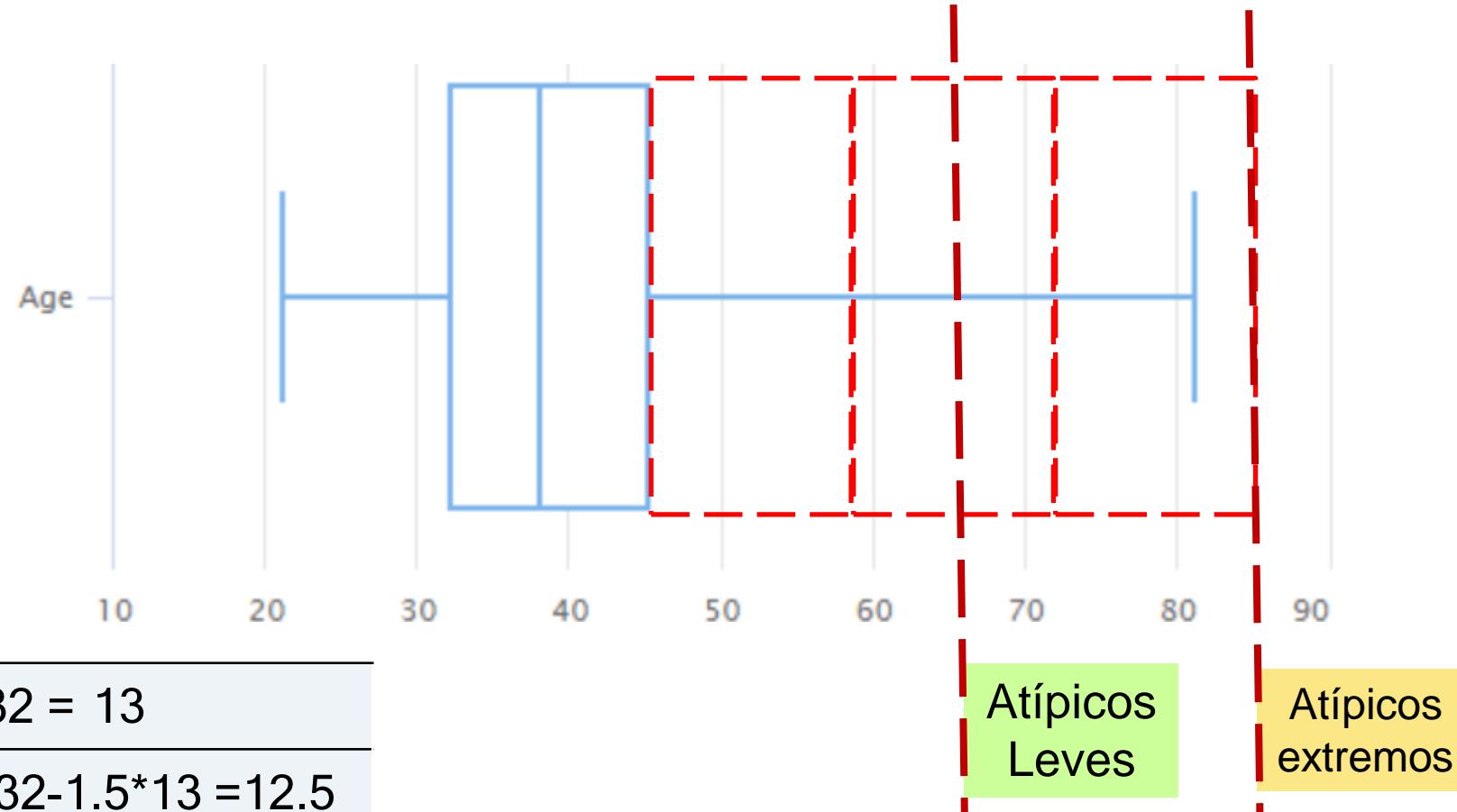
RIC	$Q3 - Q1 = 45 - 32 = 13$
Lim.Inf	$Q1 - 1.5 * RIC = 32 - 1.5 * 13 = 12.5$
Lim.Sup	$Q3 + 1.5 * RIC = 45 + 1.5 * 13 = 64.5$

Hay valores fuera
de rango?

Diagrama de caja simple

□ Atributo AGE

Minimo	21
Q1	32
Q2	38
Q3	45
Maximo	81



Valor atípico o fuera de rango

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

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

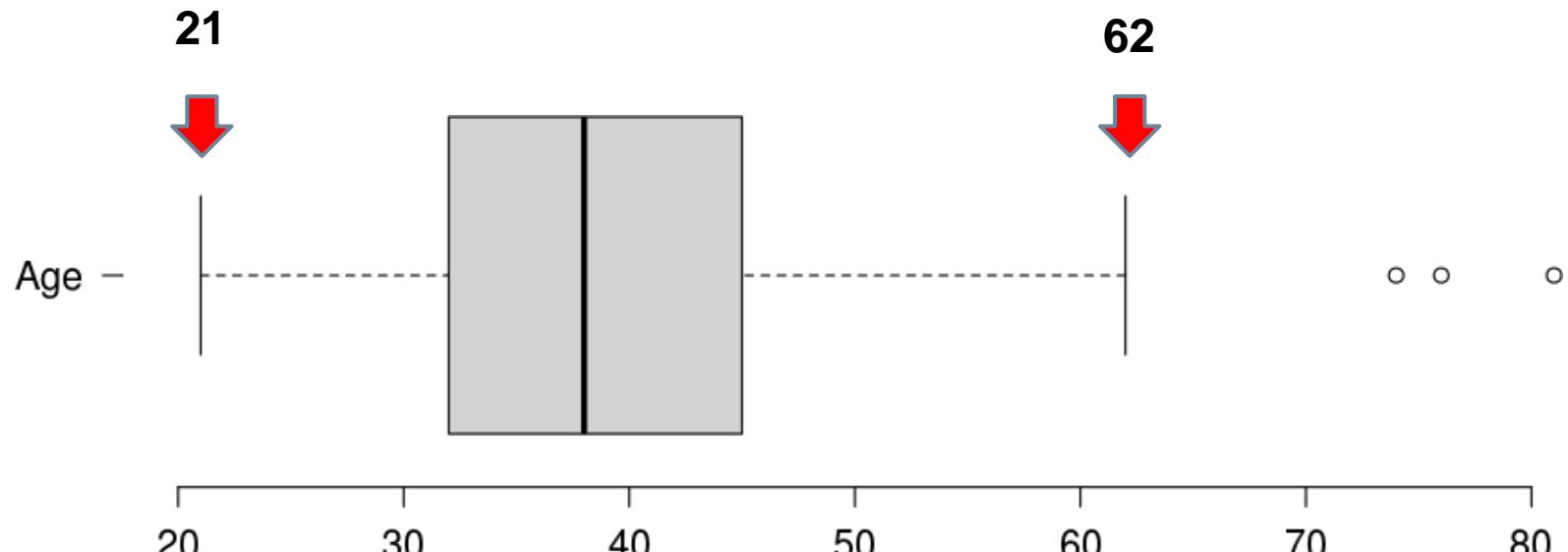
serán considerados **valores fuera de rango leves**.

- Los valores de la muestra inferiores a $Q1 - 3 \cdot RIC$ o superiores a $Q3 + 3 \cdot RIC$ serán considerados **valores fuera de rango extremos**.

Diagrama de caja de Tukey

□ Atributo AGE

Minimo	21
Q1	32
Q2	38
Q3	45
Maximo	81



RIC	13
Lim.Inf	$32 - 1.5 * 13 = 12.5$
Lim.Sup	$45 + 1.5 * 13 = 64.5$

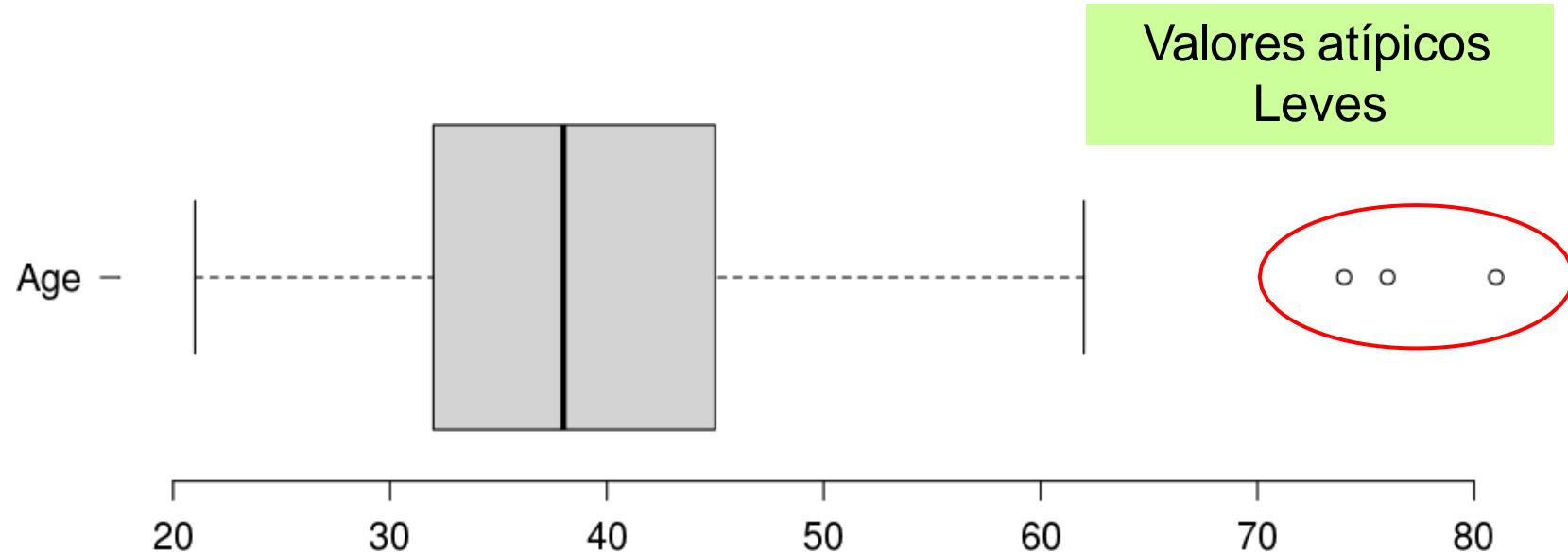
Los bigotes indican el rango de los valores de la muestra comprendidos en el intervalo

$$[Q1 - 1.5 * RIC ; Q3 + 1.5 * RIC] = [12.5 ; 64.5]$$

Diagrama de caja de Tukey

□ Atributo AGE

Minimo	21
Lim.Bigote Inf.	21
Q1	32
Q2	38
Q3	45
Lim.Bigote Sup	62
Maximo	81



- Los valores de AGE que pertenezcan a $[-7; 12.5)$ o $(64.5; 84]$ se considerarán **atípicos leves**.
- Los valores del atributo AGE inferiores a -7 o superiores a 84 se considerarán **atípicos extremos**.



Views:

Design

Results

Turbo Prep

More ▾

Find data, operators...etc



All Studio ▾

Result History

ExampleSet (Read CSV)



Data



Statistics



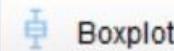
Visualizations



Annotations

Plot 1

Plot type



Value columns

Age

Group by



Plot style ➤

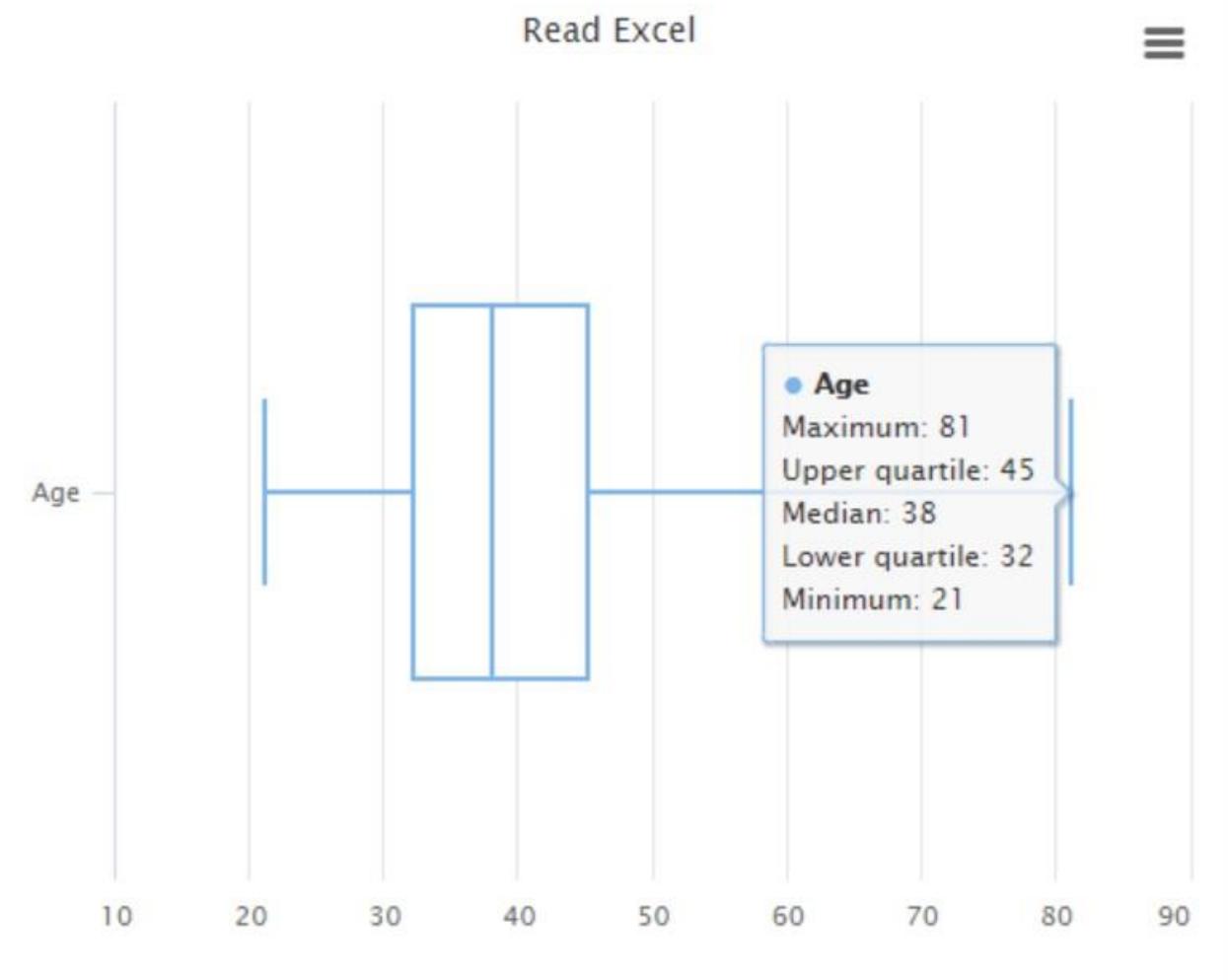
Add new plot

General



Invert chart

General style ➤

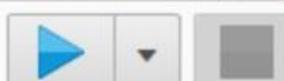


Repository

+ Import Data



- ▶ Training Resources (conn)
- ▶ Samples
- ▶ Community Samples (cor)
- ▶ DB
- ▶ ATAQUE_REDES (Laura)
- ▶ Local Repository (Laura)
- ▶ MBBS 2018 (Laura)
- ▶ MD (postgrado) (Laura)
- ▶ MD - PRACTICAS (Laura)
- ▶ MD_Educacion (Laura)
- ▶ MIDUSI (profesor)
- ▶ MINERIA (Laura)
- ▶ Mineria2018 (Laura)
- ▶ Representacion2018 (Lau
- ▶ Tesina_AR (Laura)



Views:

Design

Results

Turbo Prep

More ▾

Find data, operators...etc



All Studio ▾

Result History

ExampleSet (Read CSV)



Data



Statistics



Visualizations



Annotations

Plot

Plot 1

Plot type

Boxplot

Value column

Age

Group by

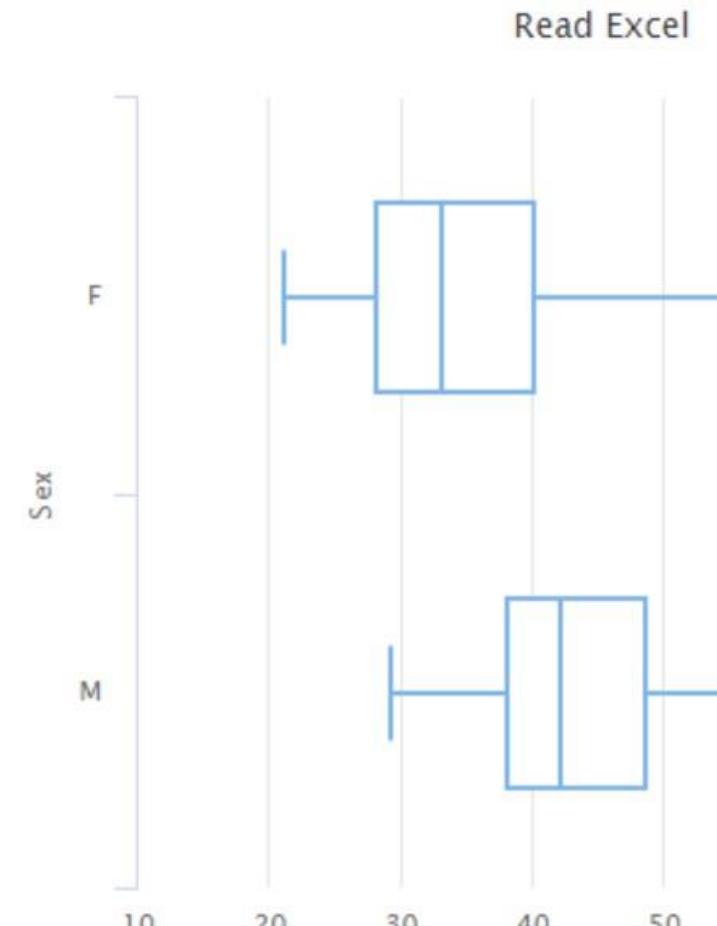
Sex

Plot style ➤

Add new plot

General

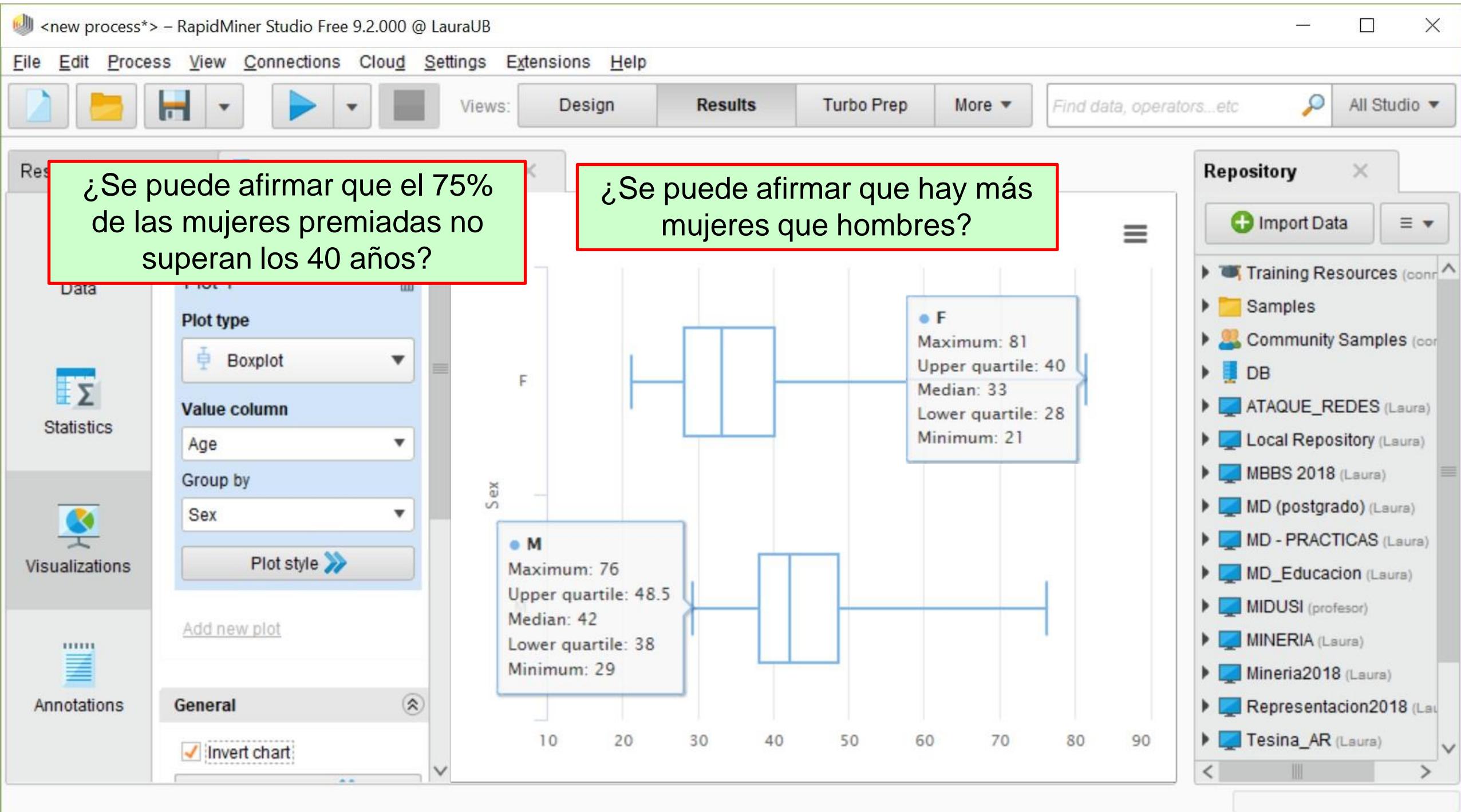
Invert chart



Repository

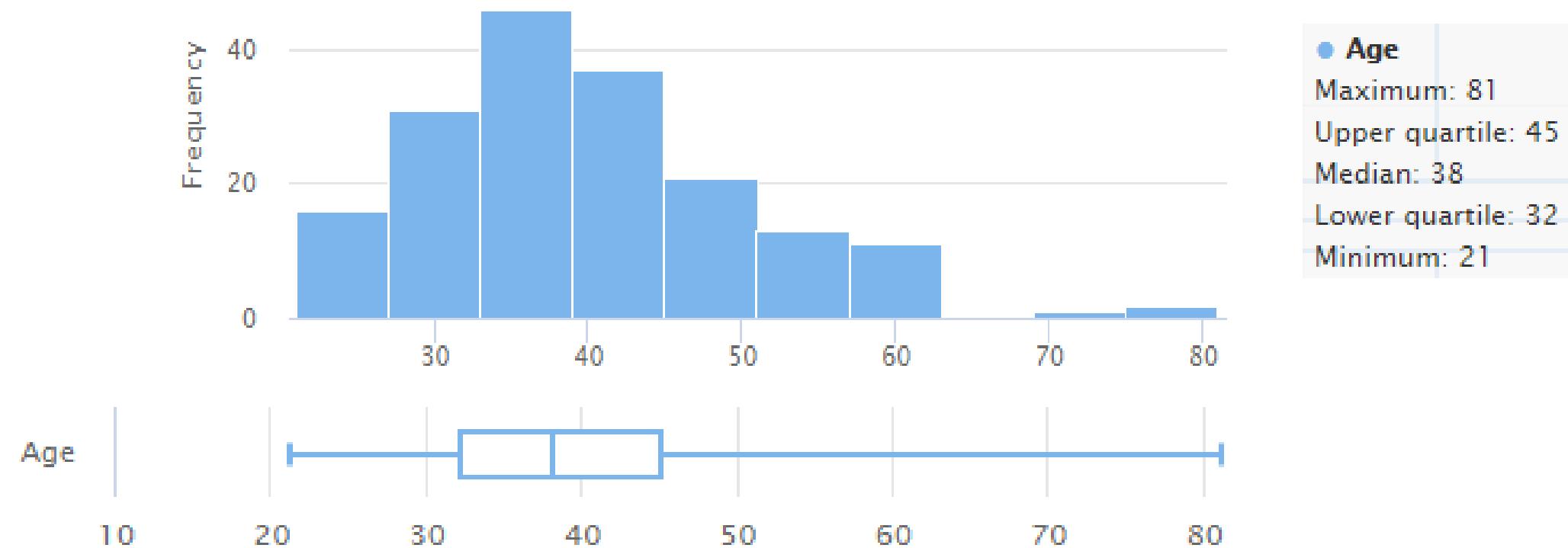
+ Import Data

- ▶ Training Resources (conn)
- ▶ Samples
- ▶ Community Samples (cor)
- ▶ DB
- ▶ ATAQUE_REDES (Laura)
- ▶ Local Repository (Laura)
- ▶ MBBS 2018 (Laura)
- ▶ MD (postgrado) (Laura)
- ▶ MD - PRACTICAS (Laura)
- ▶ MD_Educacion (Laura)
- ▶ MIDUSI (profesor)
- ▶ MINERIA (Laura)
- ▶ Mineria2018 (Laura)
- ▶ Representacion2018 (Lau
- ▶ Tesina_AR (Laura)



Histograma y diagrama de caja simple

(Atributo AGE archivo PREMIOSCSV)



Histograma y diagrama de caja de Tukey

(Atributo AGE del archivo PREMIOS.CSV)

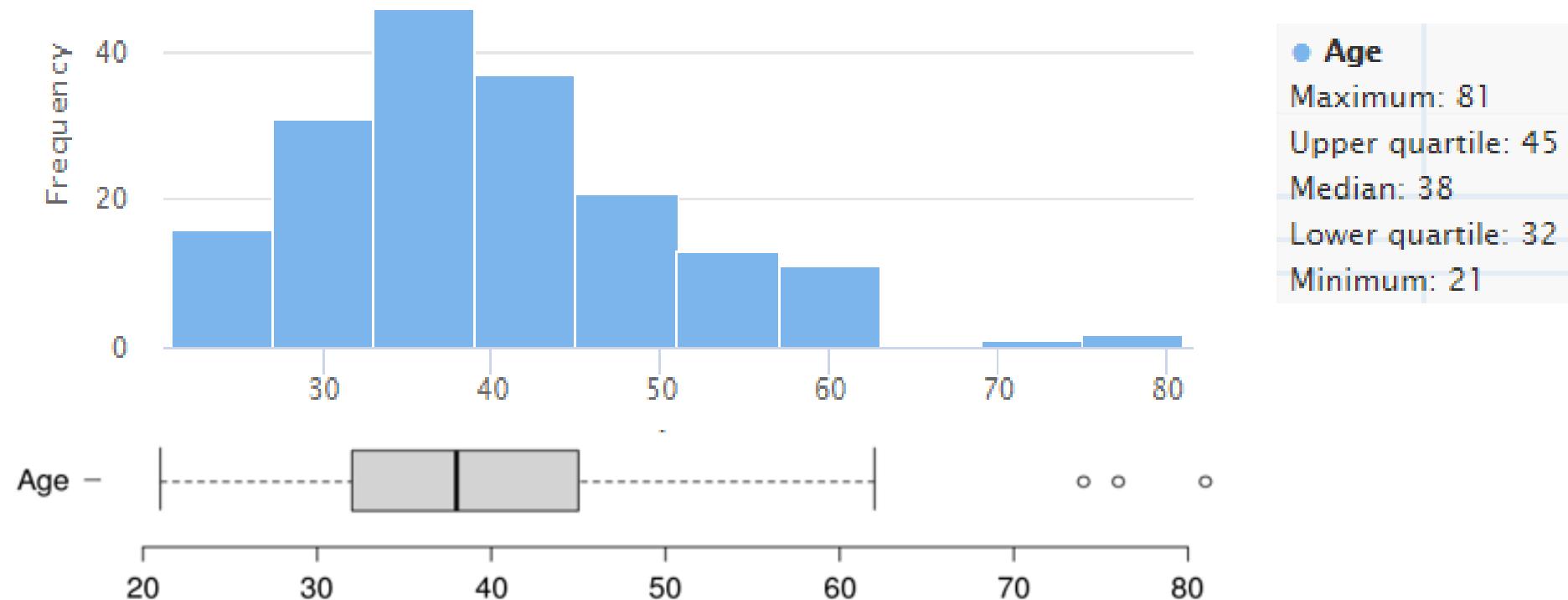


Gráfico de Torta/Dona

[File](#) [Edit](#) [Process](#) [View](#) [Connections](#) [Settings](#) [Extensions](#) [Help](#)

Views:

Design

Results

Turbo Prep



Find data, operators...etc



All Studio

Result History

ExampleSet (Read CSV)

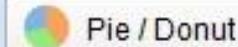


Plot

Plot 1



Plot type



Value column

genre1

 Aggregate data

Group by

genre1

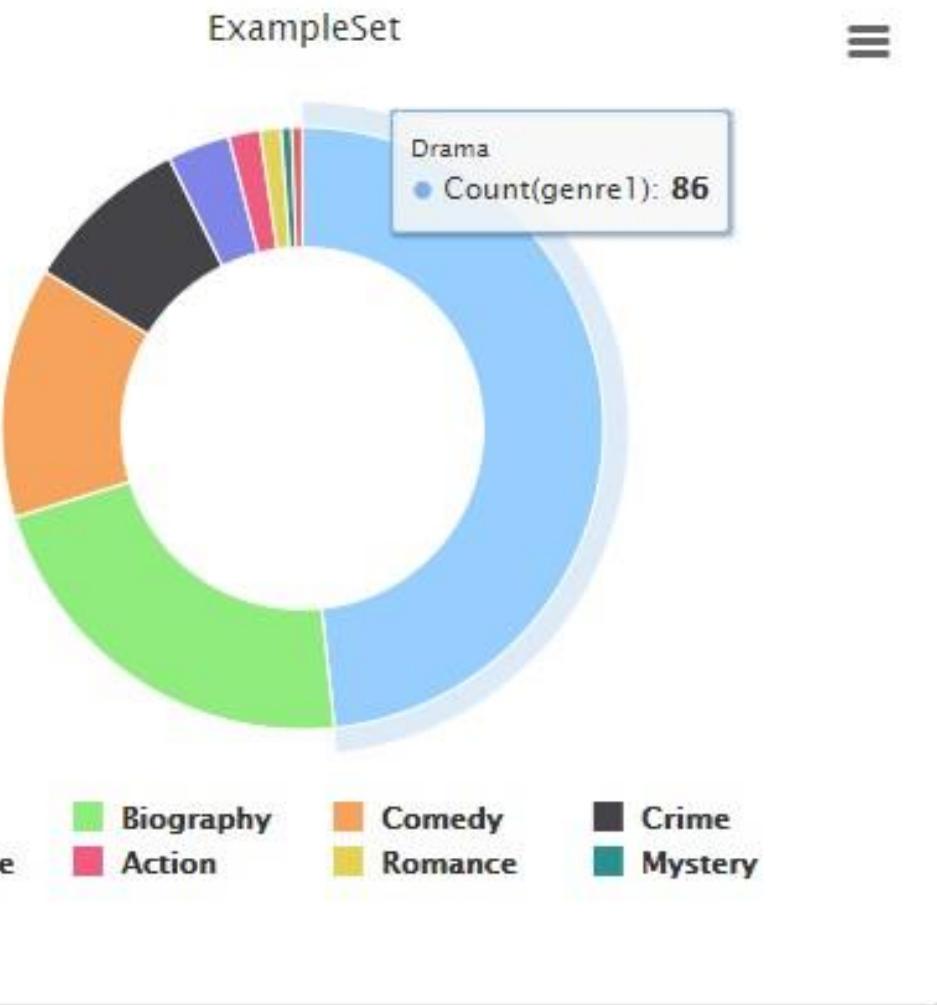


Aggregation Function

Count

 Donut

Plot style ➤



Repository

Import Data



- ▶ Training Resources (conn)
- ▶ Samples
- ▶ Community Samples (conn)
- ▶ Local Repository (Legacy)
- ▶ MD-Chilecito (Local)
- ▶ MIDUSI (Legacy)
- ▶ Representacion-Parte2 (L)
- ▶ RN-UA (Local)
- ▶ SI-USM (Local)
- ▶ DB (Legacy)

Diagrama de dispersión

[File](#) [Edit](#) [Process](#) [View](#) [Connections](#) [Settings](#) [Extensions](#) [Help](#)

Views:

Design

Results

Turbo Prep

Find data, operators...etc



All Studio

Result History

ExampleSet (Read CSV)



Data



Statistics



Visualizations



Annotations

Plot

Plot 1

Plot type

Scatter / Bubble

X-Axis column

duration

Value column

nominations

Color

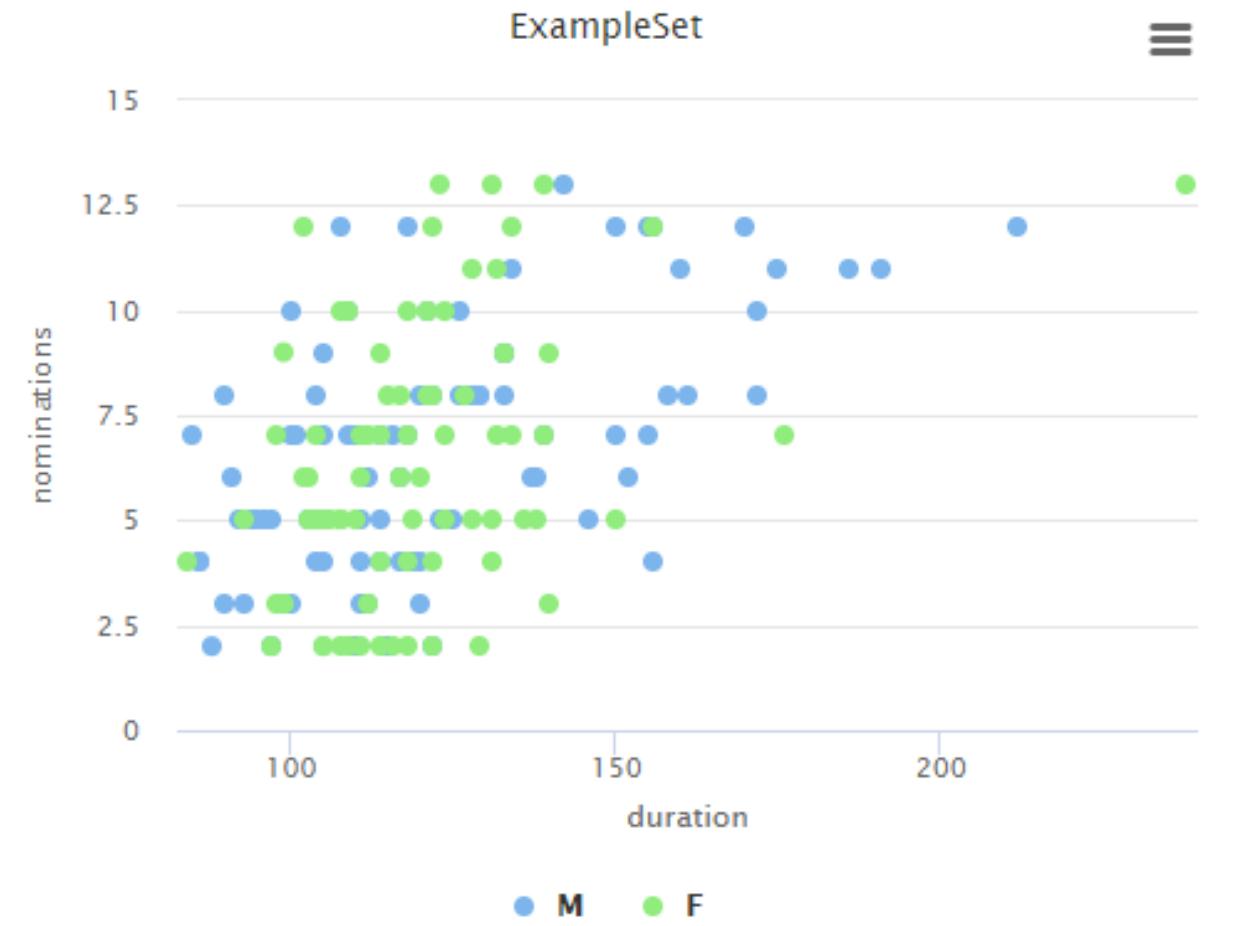
Sex

Size

-

Jitter

Regression interpolation

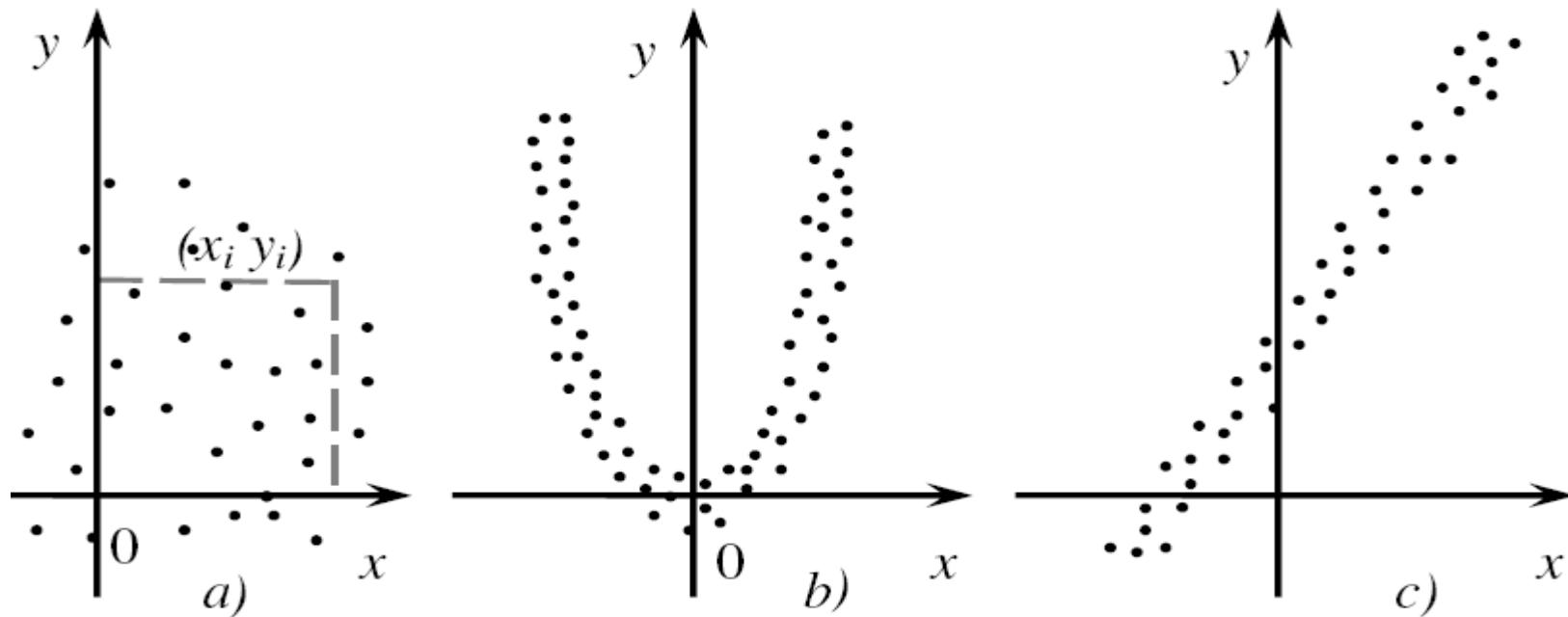


Repository

- [Import Data](#)
- Training Resources (connection)
- Samples
- Community Samples (connection)
- Local Repository (Legacy)
- MD-Chilecito (Local)
- MIDUSI (Legacy)
- Representacion-Parte2 (Local)
- RN-UA (Local)
- SI-USM (Local)
- DB (Legacy)

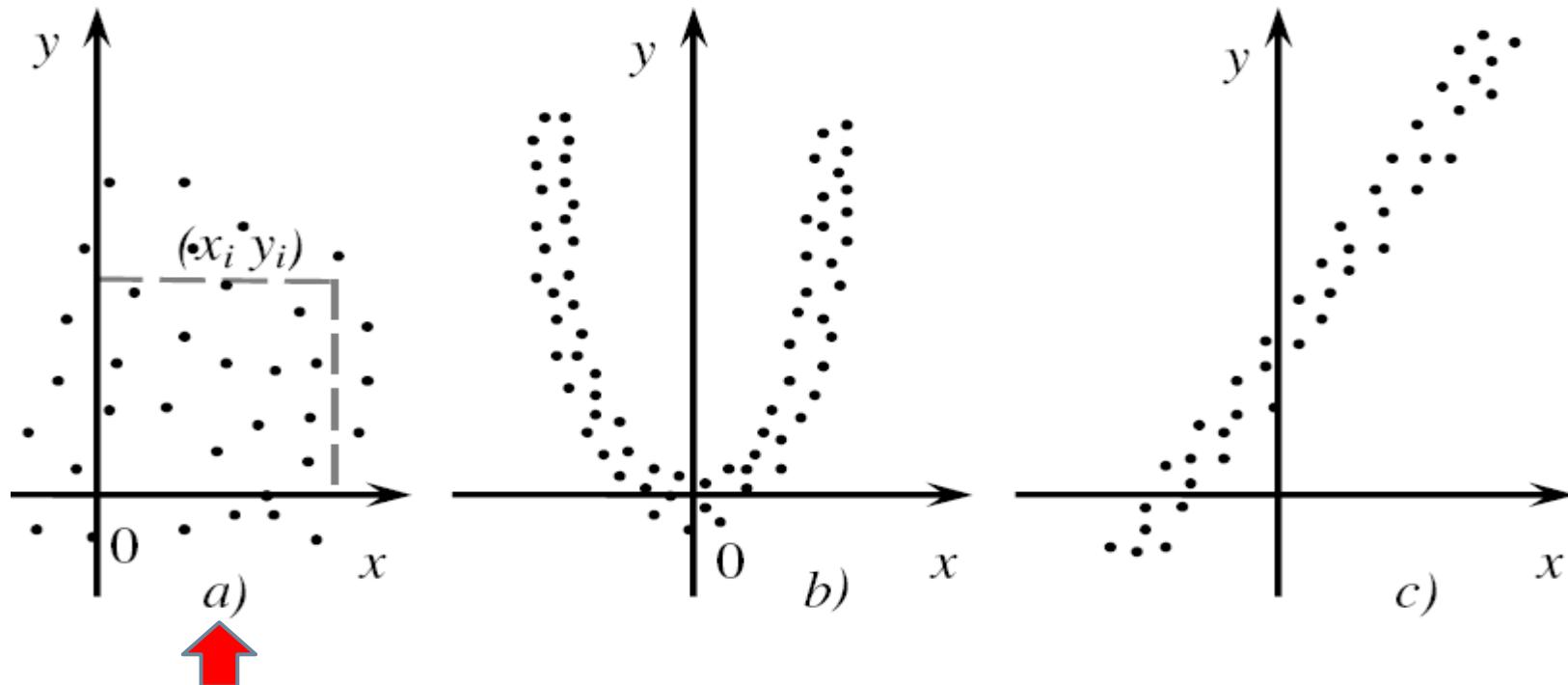
Diagramas de Dispersion

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



Diagramas de Dispersion

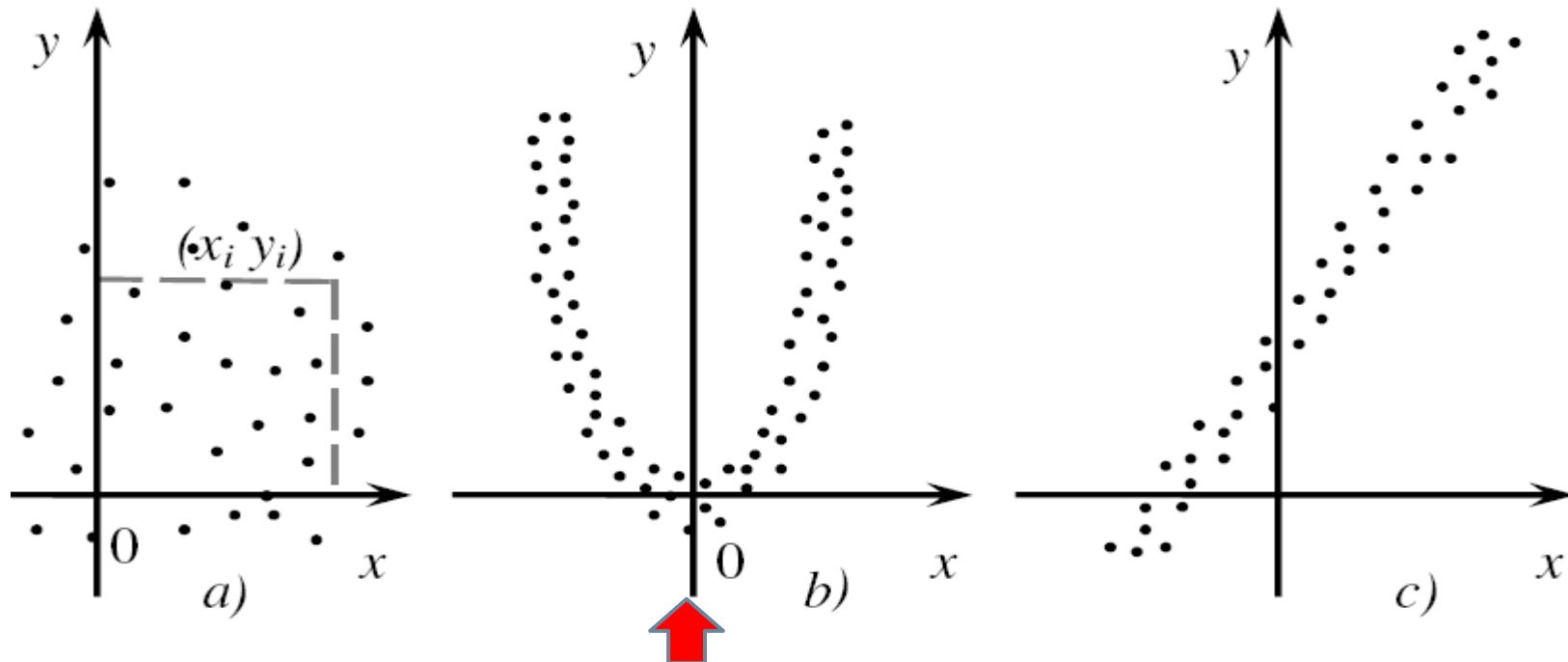
- Consiste en dibujar pares de valores (x_i, y_j) medidos de la v.a. (X, Y) en un sistema de coordenadas



Entre X e Y no hay ninguna relación funcional

Diagramas de Dispersion

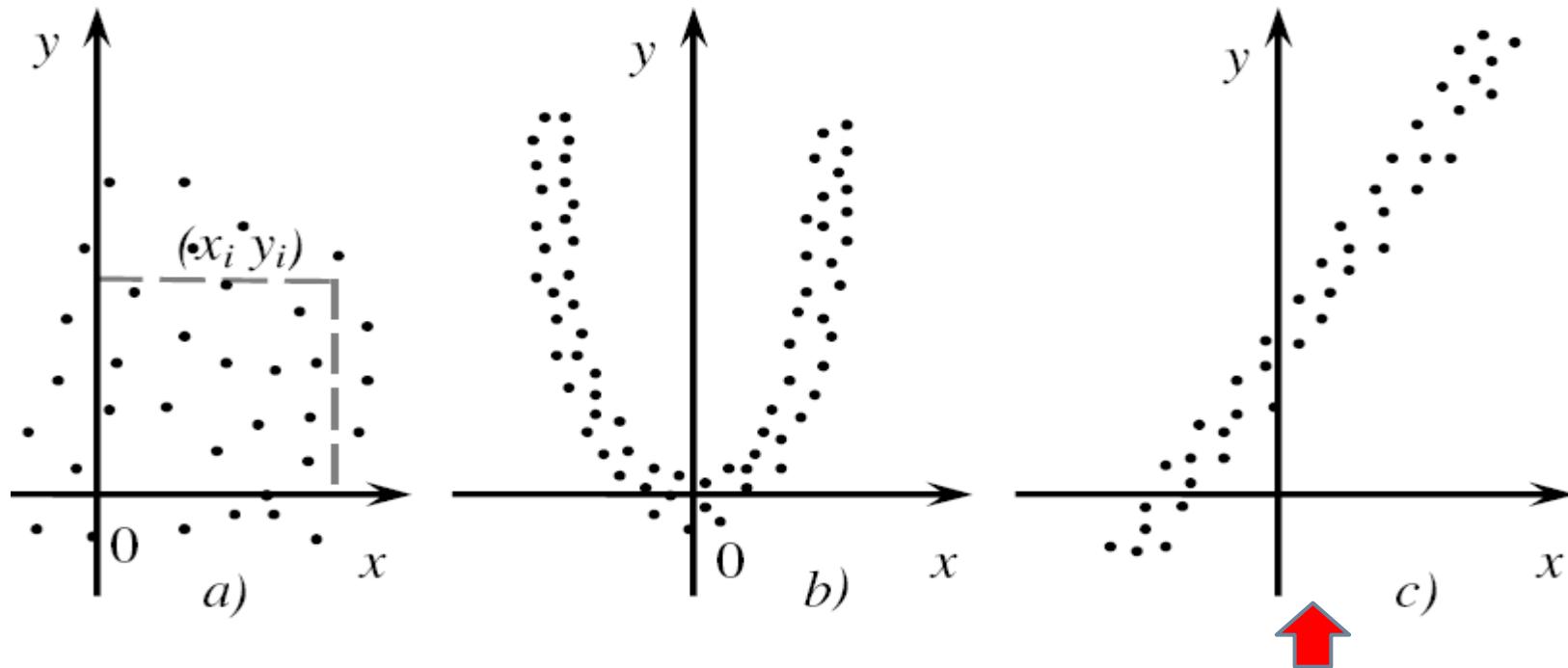
- Consiste en dibujar pares de valores (x_i, y_j) medidos de la v.a. (X, Y) en un sistema de coordenadas



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

Diagramas de Dispersion

- 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

Relación entre atributos numéricos

- 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**.

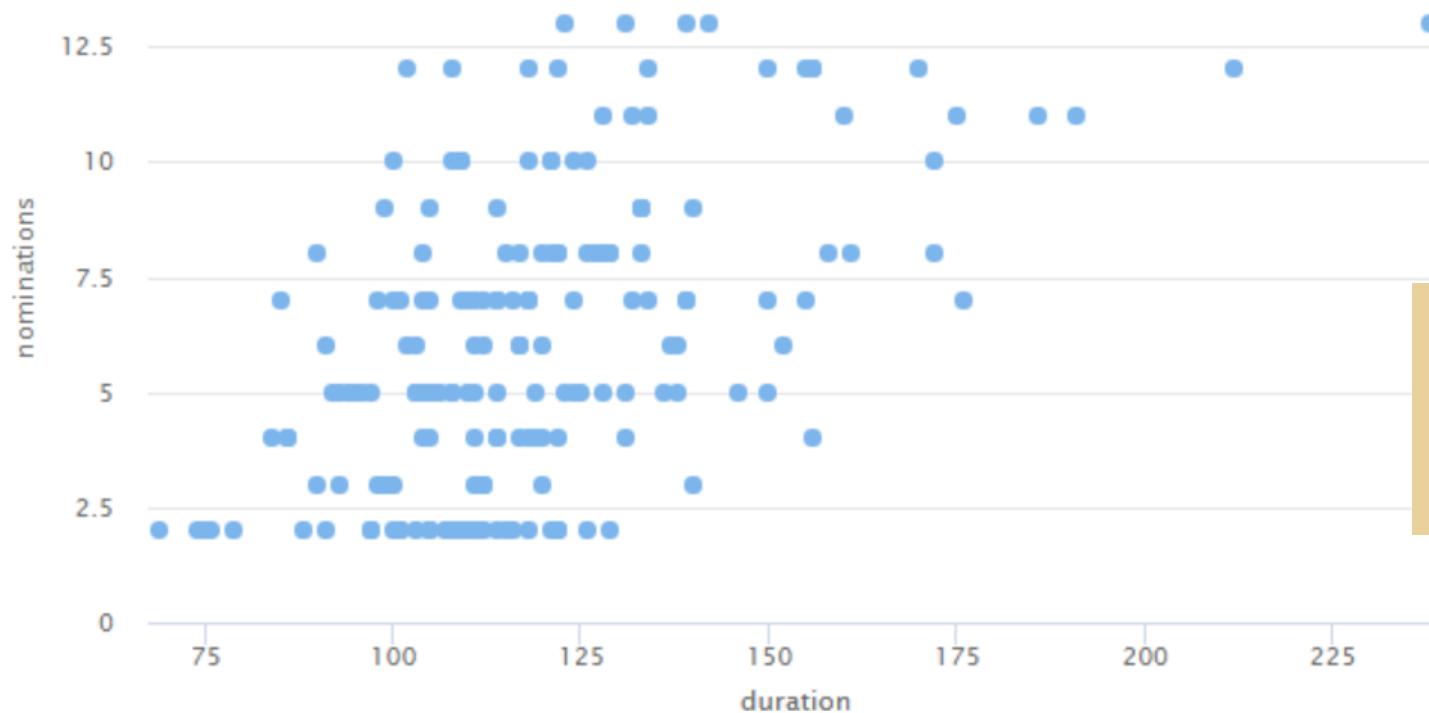


Diagrama de dispersión
entre DURATION y
NOMINATIONS

Coeficiente de correlación lineal

- Dados dos atributos X e Y el coeficiente de correlación lineal entre ellos se calcula de la siguiente forma

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \ \sigma_Y}$$

siendo $\text{Cov}(X, Y)$ la covarianza entre X e Y y σ_X y σ_Y los desvíos de cada variable.

Covarianza y desvío estándar

- Dadas dos variables X y Y

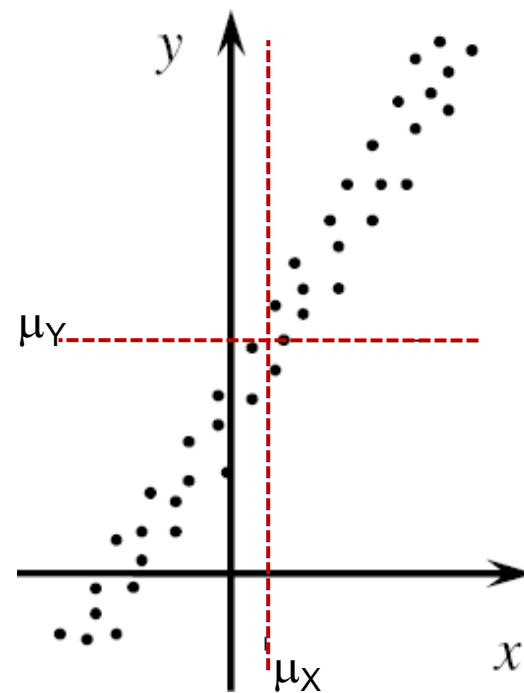
$$Cov(X, Y) = \left[\sum_{i=1}^N (x_i - \mu_X)(y_i - \mu_Y) \right] / N$$

$$\sigma_X = \sqrt{\left[\sum_{I=1}^N (x_i - \mu_X)^2 \right] / N}$$

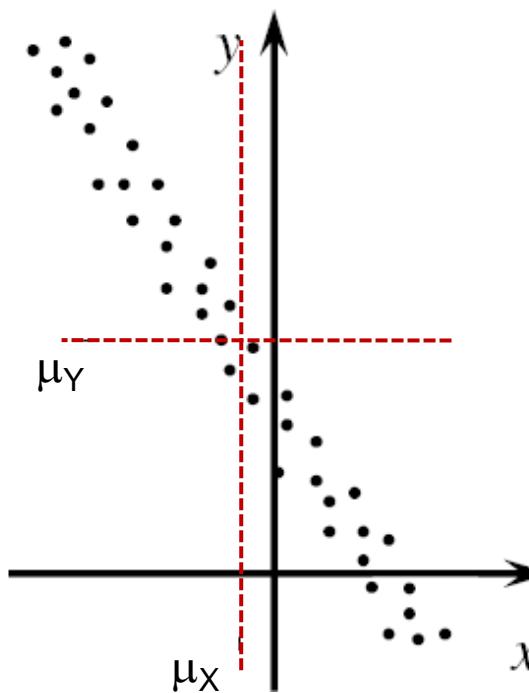
Covarianza

$$Cov(X, Y) = \left[\frac{1}{N} \sum_{i=1}^N (x_i - \mu_X)(y_i - \mu_Y) \right] / N$$

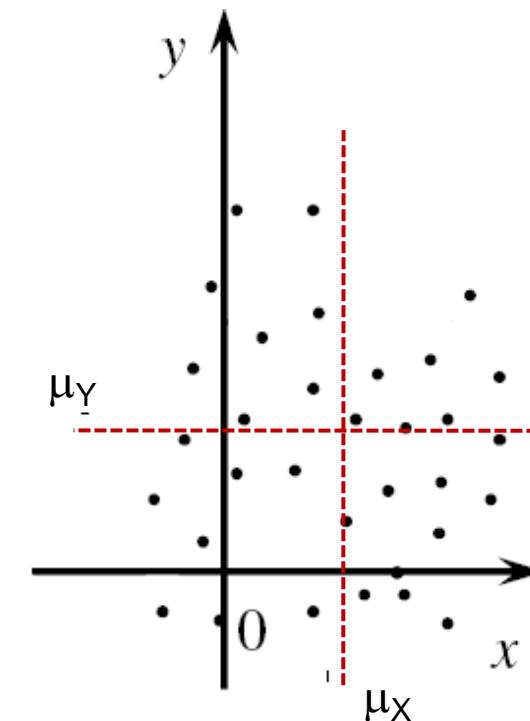
- La covarianza es un valor que indica el grado de variación conjunta de dos variables aleatorias respecto a sus medias.



Covarianza Positiva



Covarianza Negativa



Covarianza cercana a cero

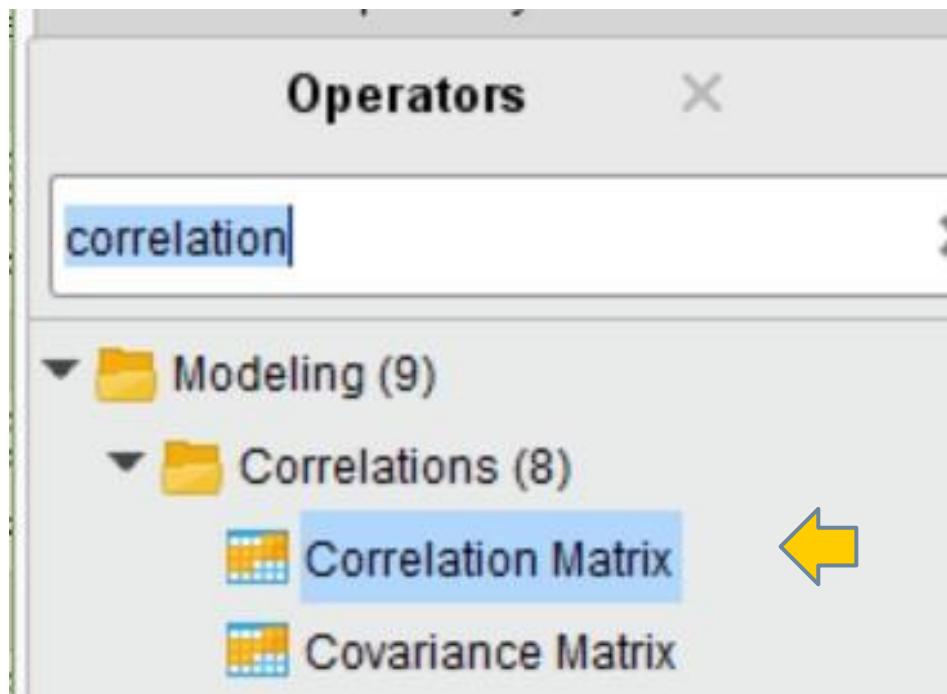
Coeficiente de correlación lineal

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.

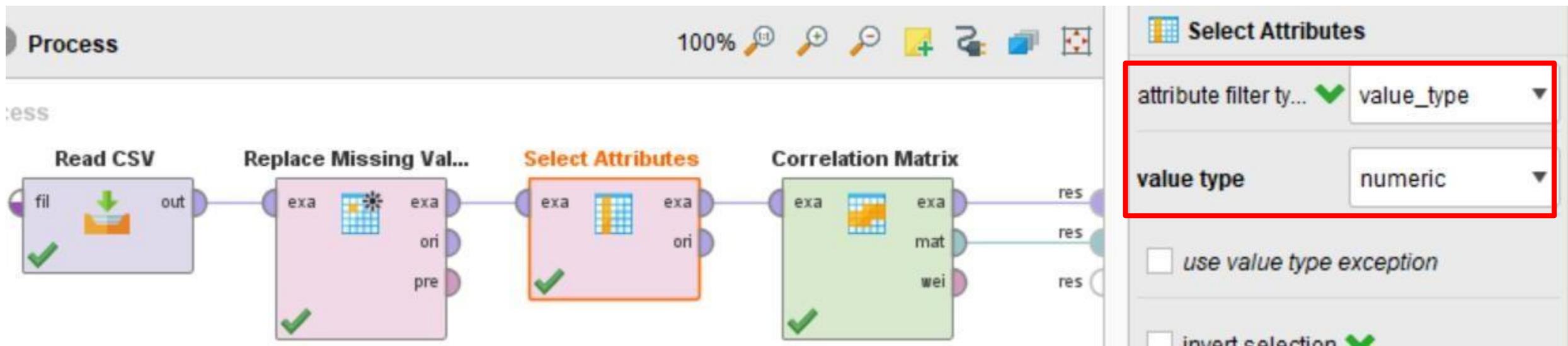
Coeficiente de correlación lineal entre atributos

- Puede utilizarse el operador **Correlation Matrix** para calcular la matriz de correlación. Recuerde que la métrica sólo aplica entre atributos numéricos.



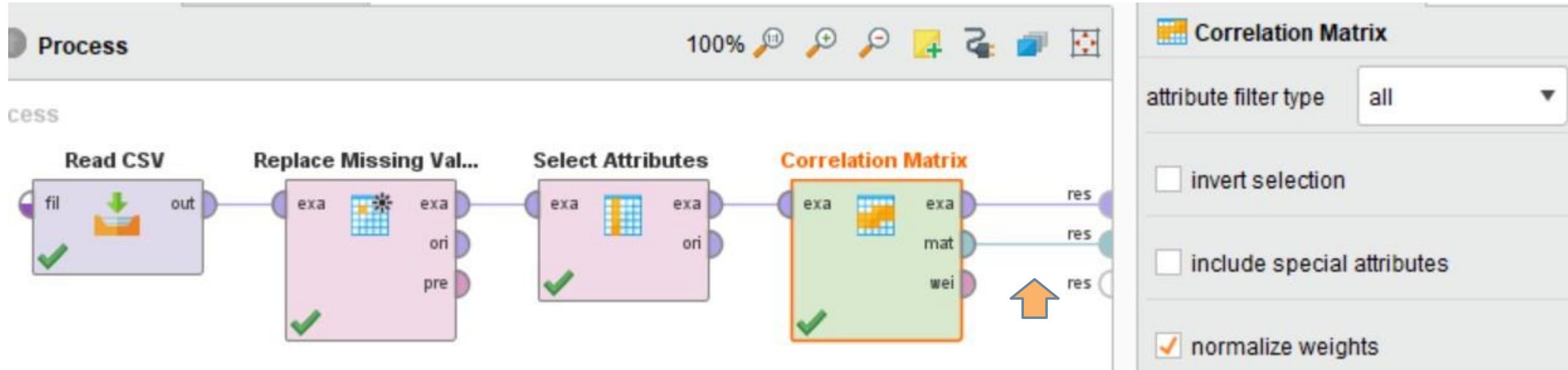
Matriz de Correlación

- Para reducir el tamaño de la matriz a construir se seleccionaron previamente los atributos numéricos



Matriz de Correlación

- Note que debe conectar la salida correspondiente a la matriz de correlación



Matriz de correlación

□ Qué significa?

Attributes	Year	Age	nominat...	rating	duration
Year	1	0.148	0.077	0.180	0.217
Age	0.148	1	0.017	-0.052	-0.080
nominations	0.077	0.017	1	0.477	0.536
rating	0.180	-0.052	0.477	1	0.424
duration	0.217	-0.080	0.536	0.424	1

Para obtener esta matriz todos los atributos deben ser numéricos y ninguno debe estar seleccionado como label

Resumen

PROCESO DE KDD

- Etapas del proceso KDD
- MD vs otras disciplinas
 - No requiere hipótesis previa
- Tipo conocimiento
 - Predictivo y descriptivo
- Tipos de variables
 - Cuantitativas y cualitativas

ANALISIS DE DATOS

- Descripciones estadísticas
 - Medidas de tendencia central
 - Medidas de dispersión
- Gráficos
 - Histograma
 - Diagrama de barras
 - Diagr. de caja simple y de Tukey
 - Diagrama de dispersión – Coeficiente de correlación lineal

Ejercicio

- Analice la información del archivo **estudiantes.csv**
 - Indique qué tipo de gráfica puede construir con los atributos. Ejemplifique cada caso.
 - La Minería de Datos permite extraer dos tipos de conocimiento: descriptivo y predictivo. Ejemplifíquelos para el caso de los estudiantes.
 - Calcule el coeficiente de correlación lineal entre los atributos numéricos. Relacione los valores obtenidos con los diagramas de dispersión de cada par de atributos.