

MINERÍA DE DATOS

weblidi.info.unlp.edu.ar/catedras/MD_SI/

Prof. Laura Lanzarini

Plantel Docente

2



III-LIDI



- **Profesora : Dra. Laura Lanzarini**
 - Temas: Redes Neuronales y Técnicas de Optimización
 - Aplicaciones en Minería de Datos y Procesamiento de Señales.

- **JTP : Dr. Facundo Quiroga** (Investigador CIC)
 - Tema: Redes neuronales profundas aplicadas al reconocimiento de patrones.

- **Ayudante : Lic. Gastón Ríos**(Becario Doctoral UNLP)
 - Tema: Generación automática de video utilizando Deep Learning.

Bibliografía

- **Introducción a la Minería de Datos**
Hernández Orallo, Ramirez Quintana, Ferri Ramirez.
Editorial Pearson – Prentice Hall. 2004

- **Data Mining. Practical Machine Learning Tools and Techniques**
Witten, Frank, Hall
Morgan Kaufmann Publisher – Elsevier - 2017

¿Cómo Aprobar?

4

Modalidad	Asistencia requerida	Ejercicios de práctica	Trabajo Final	Evaluación	Condición Final
Promoción	NO	SI	SI	SI	PROMOCIONA si $\text{nota} \geq 6$ CURSADA APROB si $4 \leq \text{nota} < 6$
Convencional	NO	NO	NO	SI	CURSADA APROB si $\text{nota} \geq 4$ y debe rendir Final

Material del Curso

5

- Toda la información y el material del curso se publicará a través de *Ideas*

ideas.info.unlp.edu.ar

- Quienes aún no tengan acceso y se encuentren inscriptos en Guaraní deberán solicitar inscripción en el curso en *Ideas*.

- Página de la cátedra

weblidi.info.unlp.edu.ar/catedras/md_si

- Contacto: **midusi.unlp@gmail.com**

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.



¿Qué es Minería de Datos?

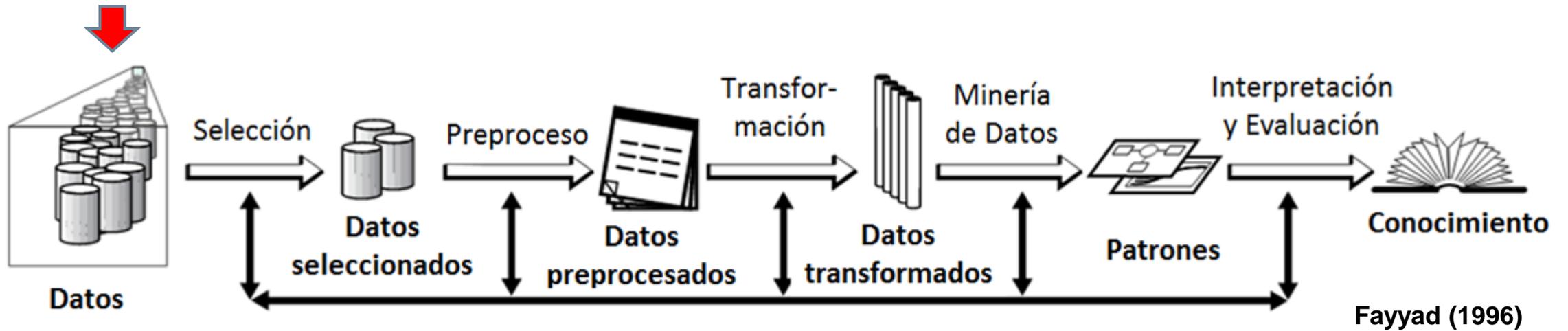
□ Es el área informática que busca descubrir **patrones** en grandes volúmenes de datos.

□ Características

- Válidos
- Novedosos
- Potencialmente útiles
- Comprensibles

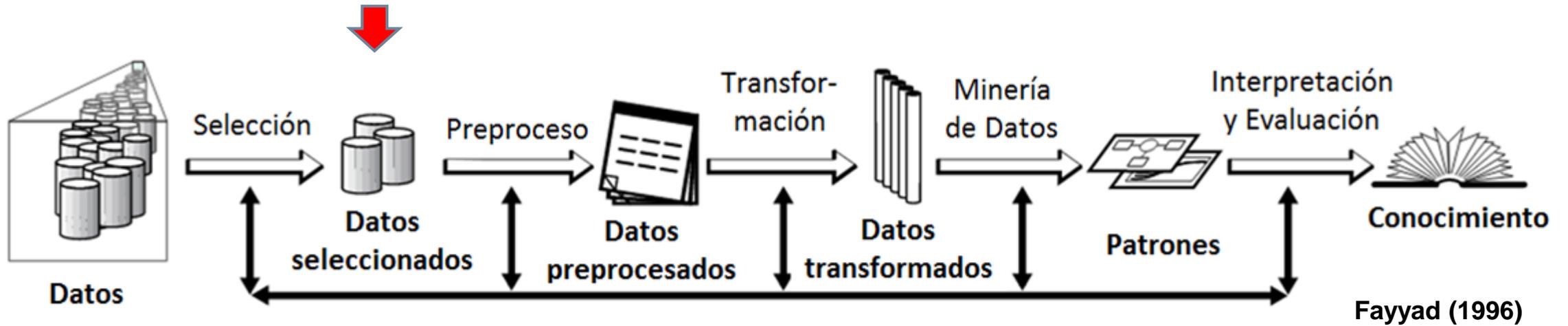


Extracción de Conocimiento (proceso KDD)



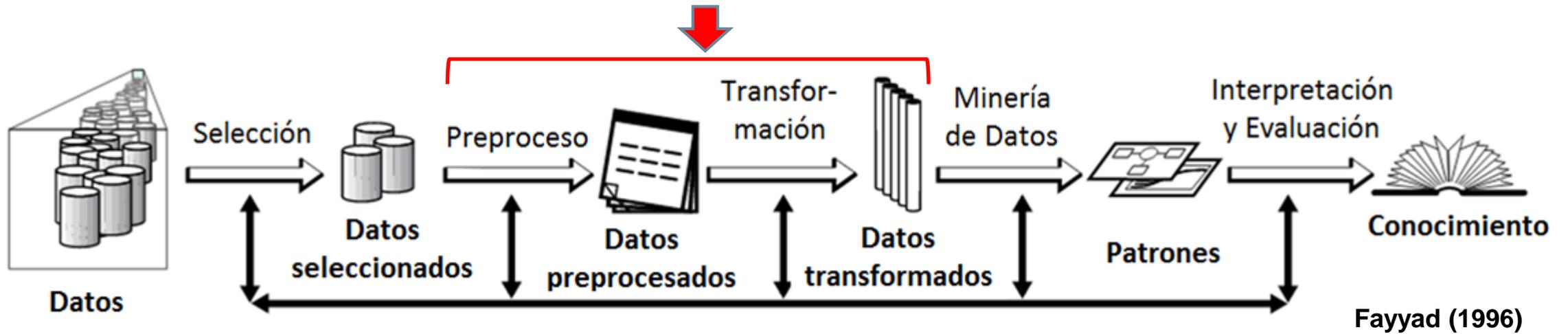
- Generalmente registrado en forma previa al proceso de KDD.
- Almacena información histórica
- No necesariamente centralizada

Extracción de Conocimiento (proceso KDD)



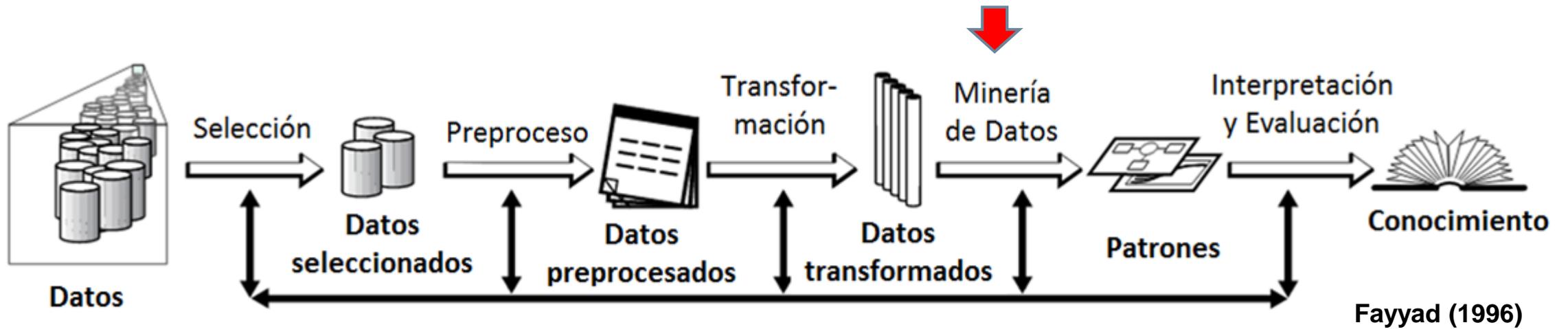
- Elegidos en base al problema
- Medidas subjetivas y objetivas

Extracción de Conocimiento (proceso KDD)

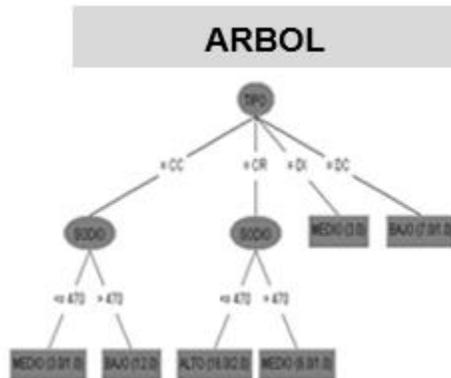


- Uniformar la notación.
- Datos faltantes
- Fuera de los rangos esperados (outliers)

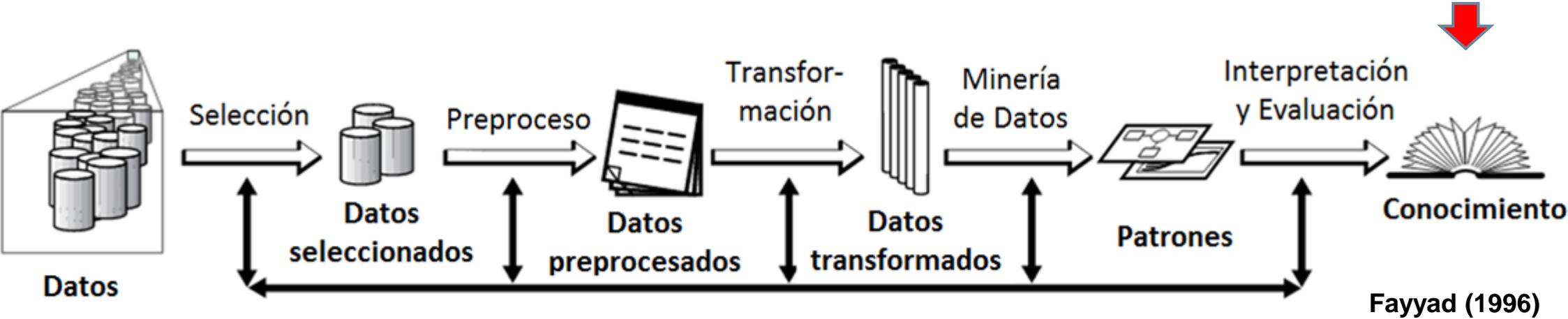
Extracción de Conocimiento (proceso KDD)



□ Técnicas de Minería de Datos



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				
12				
65				
75				
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

SI (INGRESO = APROB) entonces (RESULT=APROB)

SI (INGRESO = DESAP) AND
(FORO = NO) entonces (RESULT=DESAP)

Tipo de conocimiento a extraer

18

□ **Predictivo**

- En base 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



?

Tarea descriptiva - Aprendizaje no supervisado



AGRUPAMIENTO

Ejemplo de tarea predictiva: Prescripción de lentes

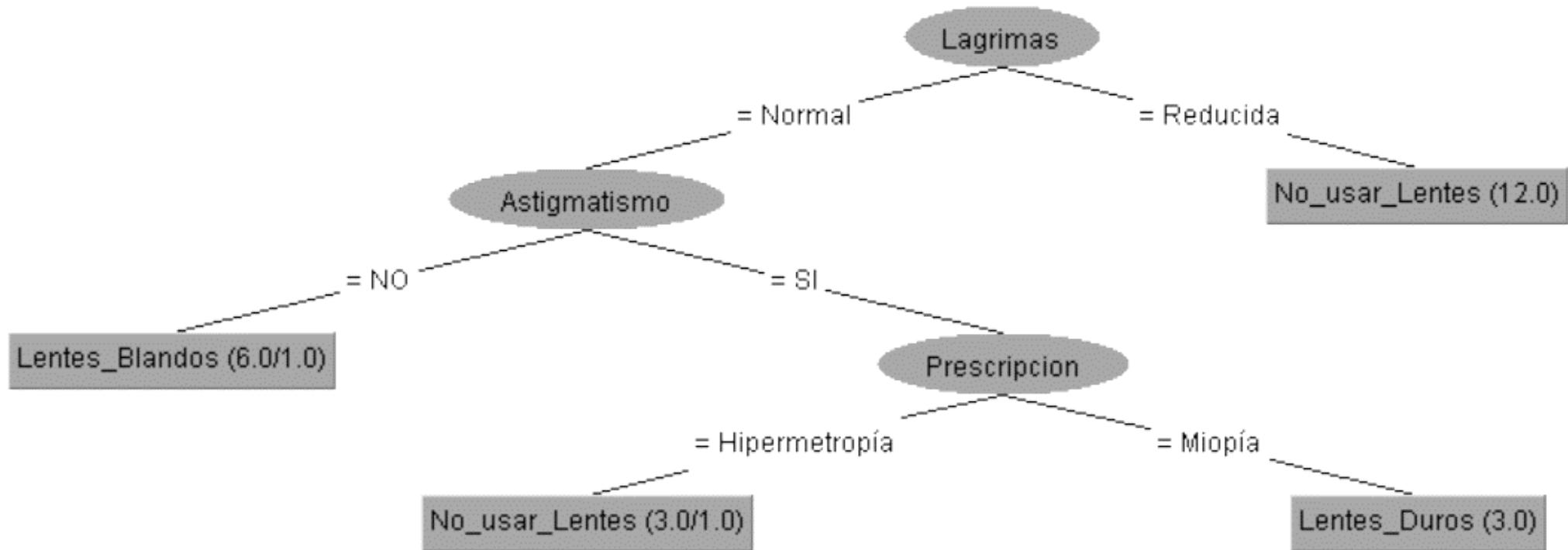
- Se dispone de la siguiente información de pacientes atendidos previamente.
 - ▣ **EDAD** del paciente: joven, pre-presbicia, presbicia
 - ▣ **PRESCRIPCION** de lentes: miope, hipermétrope
 - ▣ **ASTIGMATISMO**: si, no
 - ▣ Tasa de producción de **LAGRIMAS**: reducida, normal.
 - ▣ **DIAGNOSTICO**
 - 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	Diagnostico
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

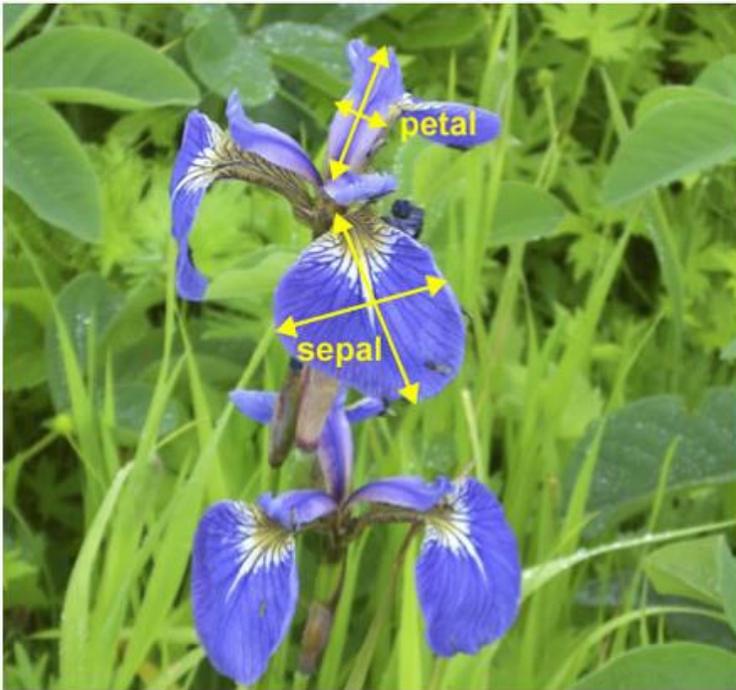
<https://archive.ics.uci.edu/ml/datasets/Lenses>

Arbol de Clasificación



Ejemplo de tarea descriptiva: Caracterización de flores

- Se dispone de información 3 tipos de flores Iris



Setosa



Versicolor



Virginica

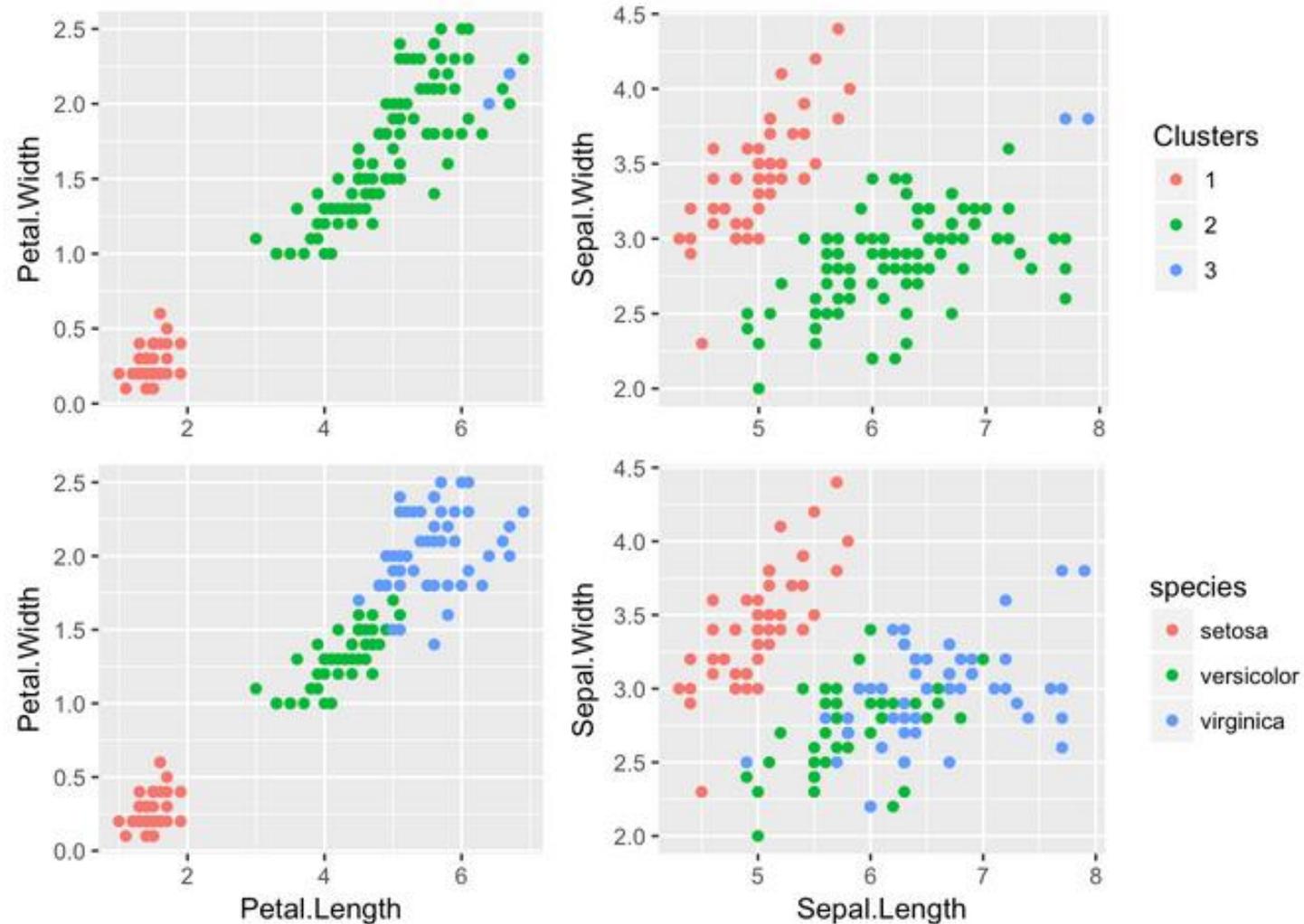
<https://archive.ics.uci.edu/ml/datasets/Iris>

Tarea descriptiva: Caracterización de flores

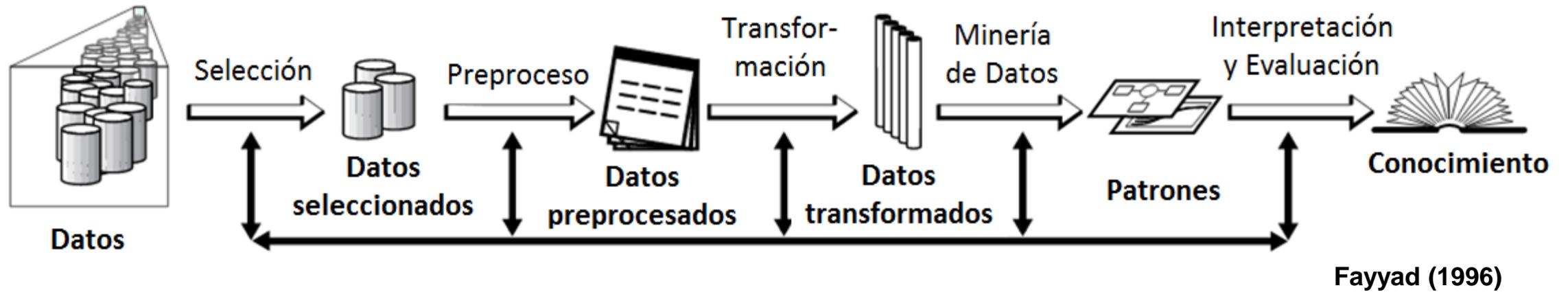
Id	sepalength	sepalwidth	petallength	petalwidth	class
1	5,1	3,5	1,4	0,2	Iris-setosa
2	4,9	3,0	1,4	0,2	Iris-setosa
...
95	5,6	2,7	4,2	1,3	Iris-versicolor
96	5,7	3,0	4,2	1,2	Iris-versicolor
97	5,7	2,9	4,2	1,3	Iris-versicolor
...
149	6,2	3,4	5,4	2,3	Iris-virginica
150	5,9	3,0	5,1	1,8	Iris-virginica

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

Tarea descriptiva: Caracterización de flores



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)

Ejemplos

- El archivo **autos-mpg.csv** contiene datos del consumo de combustible de ciertos vehículos en ciudad junto con algunas de sus características:

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

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

Ejemplo

- El archivo **autos-mpg.csv** contiene datos de 406 autos

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

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

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 : autos-mpg.csv

- Utilizaremos RapidMiner Studio para analizar la información disponible.
- Antes de comenzar, asegúrese de que dispone del archivo
autos-mpg.csv
- De no ser así, puede descargarlo del siguiente URL
weblidi.info.unlp.edu.ar/catedras/MD_SI/

<new process> - Rap

Welcome to RapidMiner Studio!

File Edit Process



Start

Recent

Learn

Operators

Repository

+ Import Data

▶ Training Resource

▶ Samples

▶ Community Sam

▶ DB

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 Repository

Import Data

- Training Resources (co)
- Samples
- Community Samples (c)
- 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 (L)

Process Help

Process

100%

Process

inp res

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

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 ▾

Repository

Operators

read

- Data Access (24)
 - Files (16)
 - Read (15)
 - Read CSV
 - Read Excel
 - Read Excel w
 - 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

Read CSV
This operator can read csv files.
Press "F3" for focus.

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

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 res

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 \(9.2.000\)](#)

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

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

Process



inp

fil

out

res

Recommended Operators

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

Conecte el operador

Parameters

Read CSV

Import Configuration Wizard...

csv file

column separat... ;

trim lines

use quotes

quotes character "

escape charact... \

[Hide advanced parameters](#)

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

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

Process

Recommended Operators

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

Parameters

Read CSV

Import Configuration Wizard...

csv file

use quotes

quotes character "

escape charact... \

[Hide advanced parameters](#)

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

Falta configurarlo

Select the data location.

Datos

Bookmarks	File Name	Size	Type	Last Modified
★ -- Last Directory	Ataques a Redes		File Folder	Jul 31, 2021
	AUTOS		File Folder	Jul 31, 2021
	adult_data.csv	3 MB	Archivo de valores separad...	Sep 27, 2018
	agaricus-lepiota.csv	364 KB	Archivo de valores separad...	Jul 27, 2020
	AIsoI.csv	1 KB	Archivo de valores separad...	Sep 12, 2018
	autos-mpg.csv	21 KB	Archivo de valores separad...	Feb 25, 2023
	balance.csv	14 KB	Archivo de valores separad...	Dec 10, 2014
	baseball.csv	22 KB	Archivo de valores separad...	Mar 29, 2022
	calabazas.csv	3 KB	Archivo de valores separad...	Mar 15, 2018
	calabazas2.csv	12 KB	Archivo de valores separad...	Sep 4, 2018
	Drug5.csv	8 KB	Archivo de valores separad...	Sep 27, 2018
	Drug5_numerico.csv	5 KB	Archivo de valores separad...	Sep 27, 2018
	Ejemplo.csv	1 KB	Archivo de valores separad...	Mar 5, 2019
	FrutasTest.csv	1 KB	Archivo de valores separad...	Oct 12, 2018
	moons_unlabeled.csv	9 KB	Archivo de valores separad...	Sep 25, 2018

autos-mpg.csv

CSV (.tsv, .csv)

← Previous → Next ✖ Cancel

Select the data location.

Datos

Bookmarks

- ★ -- Last Directory

File Name	Size	Type	Last Modified
Ataques a Redes		File Folder	Jul 31, 2021
AUTOS		File Folder	Jul 31, 2021
adult_data.csv	3 MB	Archivo de valores separad...	Sep 27, 2018
agaricus-lepiota.csv	364 KB	Archivo de valores separad...	Jul 27, 2020
AIsoI.csv	1 KB	Archivo de valores separad...	Sep 12, 2018
autos-mpg.csv	21 KB	Archivo de valores separad...	Feb 25, 2023
balance.csv	14 KB	Archivo de valores separad...	Dec 10, 2014
baseball.csv	22 KB	Archivo de valores separad...	Mar 29, 2022
calabazas.csv	3 KB	Archivo de valores separad...	Mar 15, 2018
calabazas2.csv	12 KB	Archivo de valores separad...	Sep 4, 2018
Drug5.csv	8 KB	Archivo de valores separad...	Sep 27, 2018
Drug5_numerico.csv	5 KB	Archivo de valores separad...	Sep 27, 2018
Ejemplo.csv	1 KB	Archivo de valores separad...	Mar 5, 2019
FrutasTest.csv	1 KB	Archivo de valores separad...	Oct 12, 2018
moons_unlabeled.csv	9 KB	Archivo de valores separad...	Sep 25, 2018

autos-mpg.csv

CSV (.tsv, .csv)

← Previous **Next** → ✕ Cancel

Specify your data format

<input checked="" type="checkbox"/> Header Row	<input type="text" value="1"/>	File Encoding	<input type="text" value="windows-1252"/>	<input checked="" type="checkbox"/> Use Quotes	<input type="text" value=""/>
Start Row	<input type="text" value="1"/>	Escape Character	<input type="text" value="\"/>	<input type="checkbox"/> Trim Lines	
Column Separator	<input type="text" value="Semicolon ';'"/>	Decimal Character	<input type="text" value="."/>	<input checked="" type="checkbox"/> Skip Comments	<input type="text" value="#"/>

1	mpg	cylinders	displacement	horsepower	weight	acceleration	model_year	origin	car_name
2	18.0	ocho	3070	130.0	3504	120	70	1	chevrolet chev...
3	15.0	ocho	3500	165.0	3693	115	70	1	buick skylark ...
4	18.0	ocho	3180	150.0	3436	110	70	1	plymouth sate...
5	16.0	ocho	3040	150.0	3433	120	70	1	amc rebel sst
6	17.0	ocho	3020	140.0	3449	105	70	1	ford torino
7	15.0	ocho	4290	198.0	4341	100	70	1	ford galaxie 5...
8	14.0	ocho	4540	220.0	4354	90	70	1	chevrolet imp...
9	14.0	ocho	4400	215.0	4312	85	70	1	plymouth fury iii
10	14.0	ocho	4550	225.0	4425	100	70	1	pontiac catalina



✔ no problems.



Format your columns.

Date format

Replace errors with missing values ⓘ

	mpg <i>real</i>	cylinders <i>polynomial</i>	displacement <i>integer</i>	horsepower <i>real</i>	weight <i>integer</i>	acceleration <i>integer</i>	model_year <i>integer</i>	origin <i>integer</i>
1	18.000	ocho	3070	130.000	3504	120	70	1
2	15.000	ocho	3500	165.000	3693	115	70	1
3	18.000	ocho	3180	150.000	3436	110	70	1
4	16.000	ocho	3040	150.000	3433	120	70	1
5	17.000	ocho	3020	140.000	3449	105	70	1
6	15.000	ocho	4290	198.000	4341	100	70	1
7	14.000	ocho	4540	220.000	4354	90	70	1
8	14.000	ocho	4400	215.000	4312	85	70	1
9	14.000	ocho	4550	225.000	4425	100	70	1
10	15.000	ocho	3900	190.000	3850	85	70	1



✔ no problems.

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

Operators Repository

read

- Data Access (25)
 - Files (15)
 - Read (14)
 - Read CSV
 - Read Excel
 - Read URL
 - Read SPSS
 - Read Stata
 - Read Sparse
 - Read ARFF
 - Read XRFF
 - Read DBase
 - Read C4.5

We found "Spreadsheet Table Extraction", "SAS Connector" and 5 more results in the Marketplace. [Show me!](#)

Process Help

Process

Ejecutar

Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

Activate Wisdom of Crowds

Parameters

Read CSV

Import Configuration Wizard...

csv file

column separators

trim lines

use quotes

quotes character

escape character

skip comments

comment characters

[Hide advanced parameters](#)

[Change compatibility \(9.10.013\)](#)

Views **Design** **Results** Turbo Prep Auto Model More Find data, operators...etc All Studio

Result History ExampleSet (Read CSV)

Data
Statistics
Visualizations
Annotations

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

Open in Turbo Prep (406 / 406 examples): all

Row No.	mpg	acceleration	model_year
1	18	120	70
2	15	115	70
3	18	110	70
4	16	120	70
5	17	105	70
6	15	100	70
7	14	90	70
8	14	85	70
9	14	100	70
10	15	85	70
11	?	175	70

Repository

Import Data

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

ExampleSet (406 examples, 0 special attributes, 9 regular attributes)

Result History ExampleSet (Read CSV)

Metadatos

Name	Type	Missing	Filter (9 / 9 attributes):	Search for Attributes
	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
▼ acceleration	Integer	0	Min 80	Max 950

Statistics

Visualizations

Annotations

Repository

- Import Data
- Training Resources (conn)
- Samples
- Community Samples (cor)
- AAP22 (Local)
- Local Repository (Local)
- MD2022 (Local)
- Repre_2023 (Local)
- DB (Legacy)

Showing attributes 1 - 9

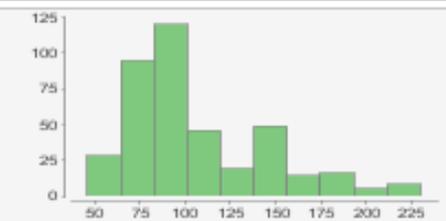
Examples: 406 Special Attributes: 0 Regular Attributes: 9

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

Result History ExampleSet (Read CSV)

Al clickear sobre el atributo muestra más información

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



Acceso al gráfico

Statistics

Visualizations

Annotations

Repository

- Import Data
- Training Resources (conn)
- Samples
- Community Samples (cor)
- AAP22 (Local)
- Local Repository (Local)
- MD2022 (Local)
- Repre_2023 (Local)
- DB (Legacy)

Si se trata de una variable cuantitativa muestra el histograma

Result History

- Data
- Statistics
- Visualizations**
- Annotations

Plot 1

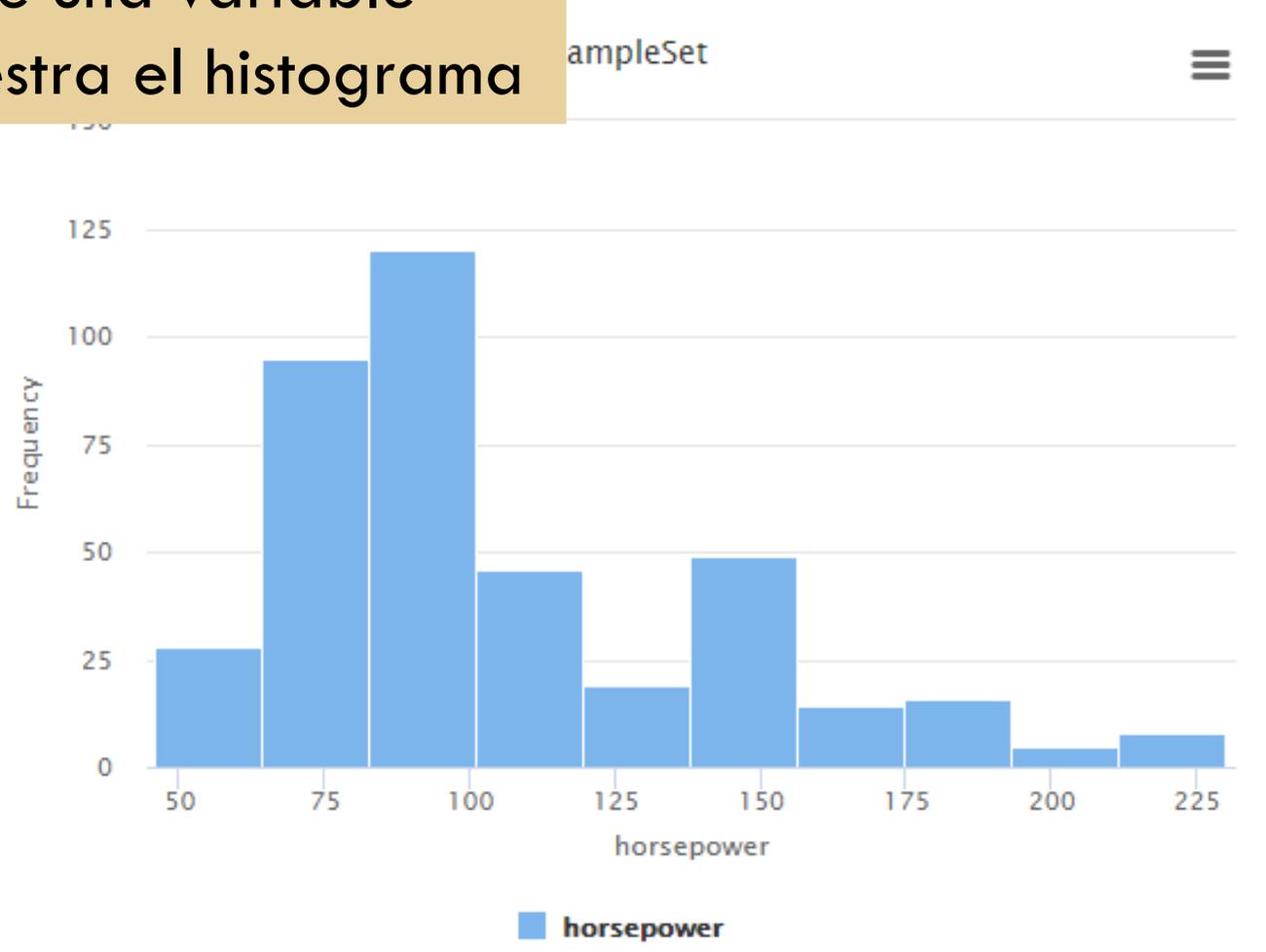
Plot type: Histogram

Value columns: horsepower

Color: -

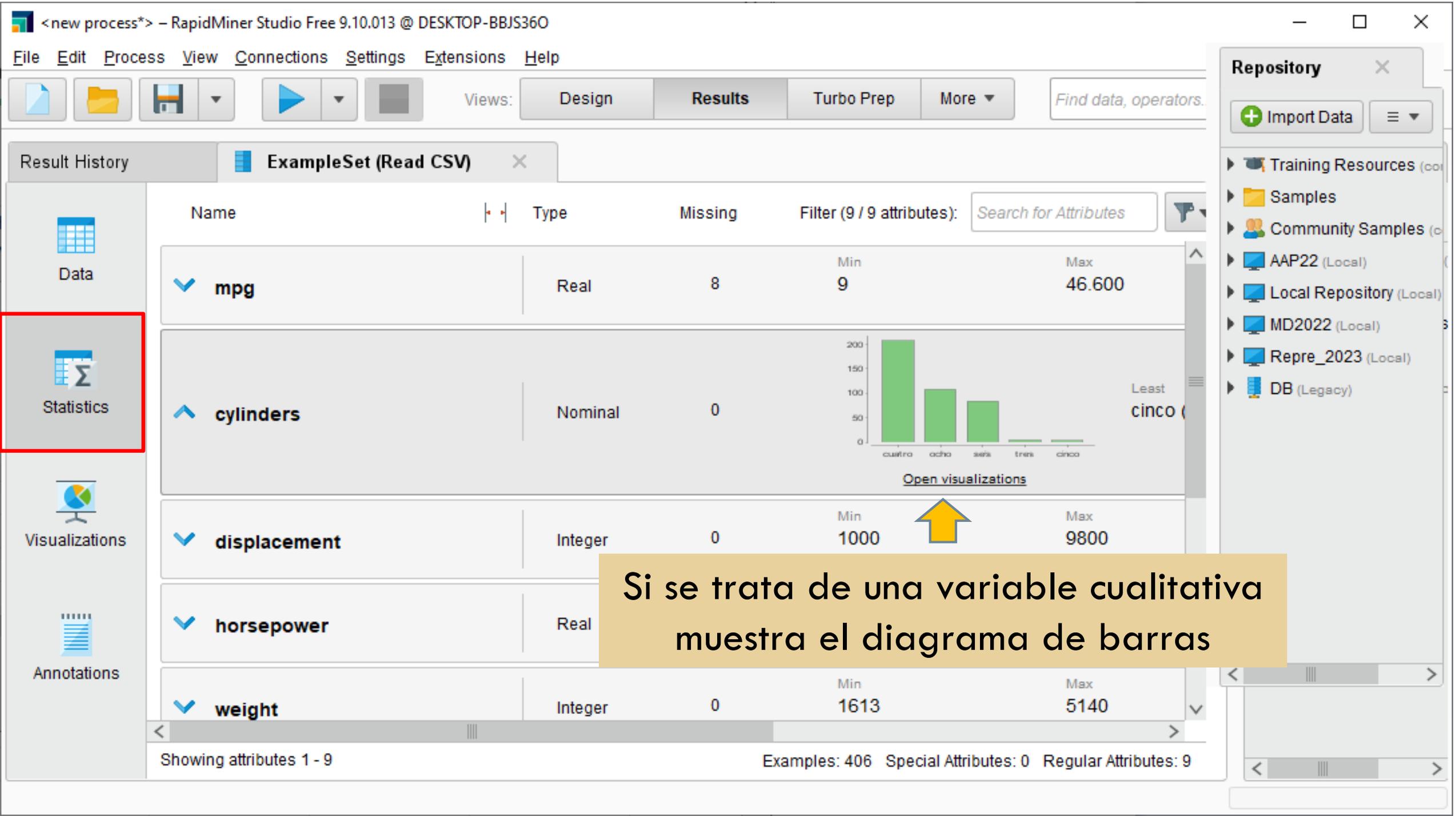
Number of Bins: 10

Plot style



Repository

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



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

Result History ExampleSet (Read CSV)

- Data
- Statistics
- Visualizations
- Annotations

Plot

Plot 1

Plot type: Bar (Column)

Value columns: cylinders

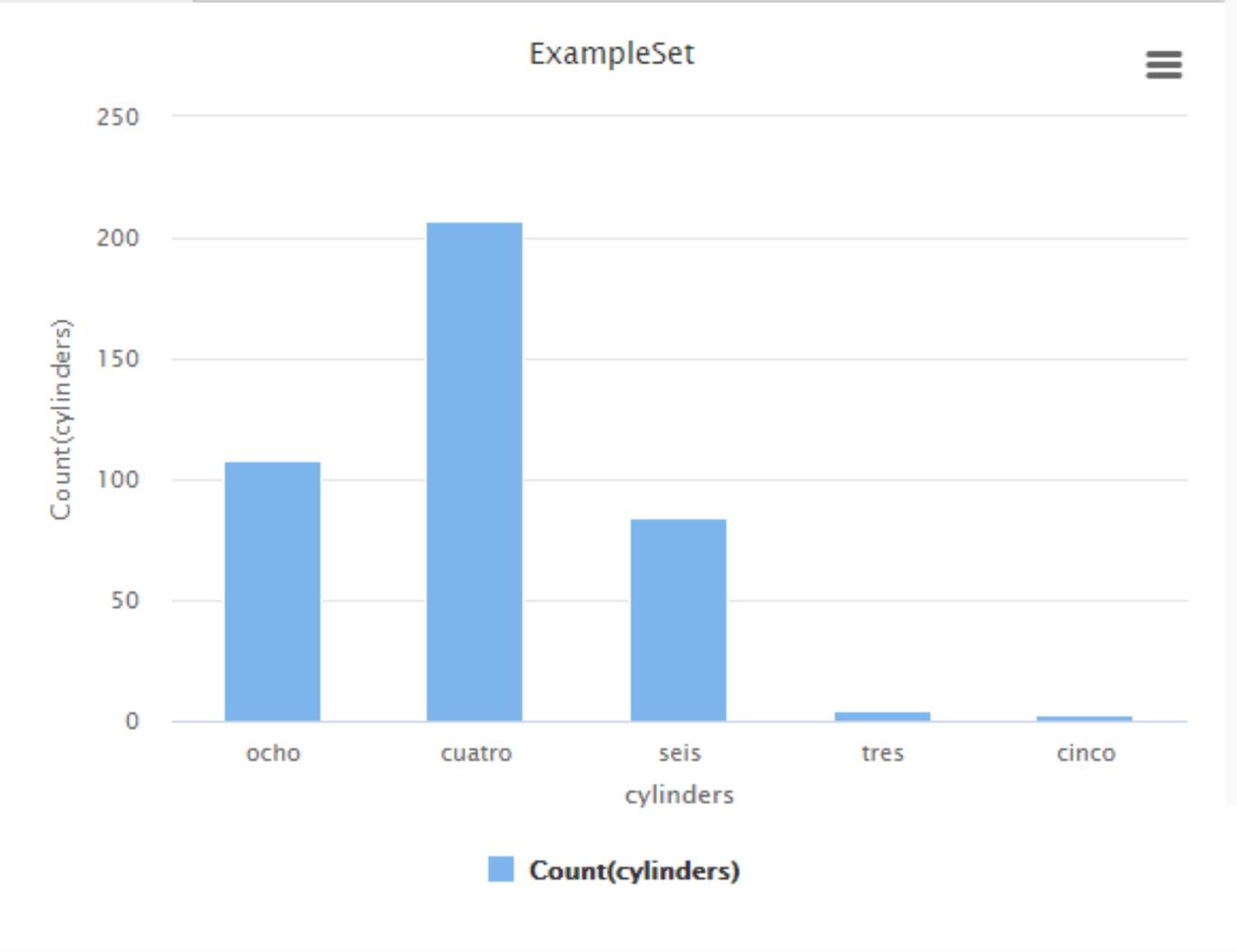
Aggregate data

Group by: cylinders

Aggregation Function: Count

Color Group: -

Stacking: No stacking



Repository

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

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

MEDIDAS DE DISPERSION

- 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

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

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

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



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

MEDIANA

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

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



$$\tilde{X} = 56$$

MEDIANA

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

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


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

MEDIANA

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

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



$$\tilde{X} = 54$$

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

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



$$\tilde{X} = \text{medio}$$

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	medio	medio	grande	grande	grande
-------	-------	-------	-------	-------	--------	--------	--------



$$\tilde{X} = \text{medio}$$

MEDIANA

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

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



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

MODA

- La moda es el valor que aparece con mayor frecuencia. Por lo tanto, puede determinarse para atributos cualitativos y cuantitativos.
- Es posible que la mayor frecuencia corresponda a varios valores diferentes, lo que da lugar a más de una MODA.
- 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 HORSEPOWER

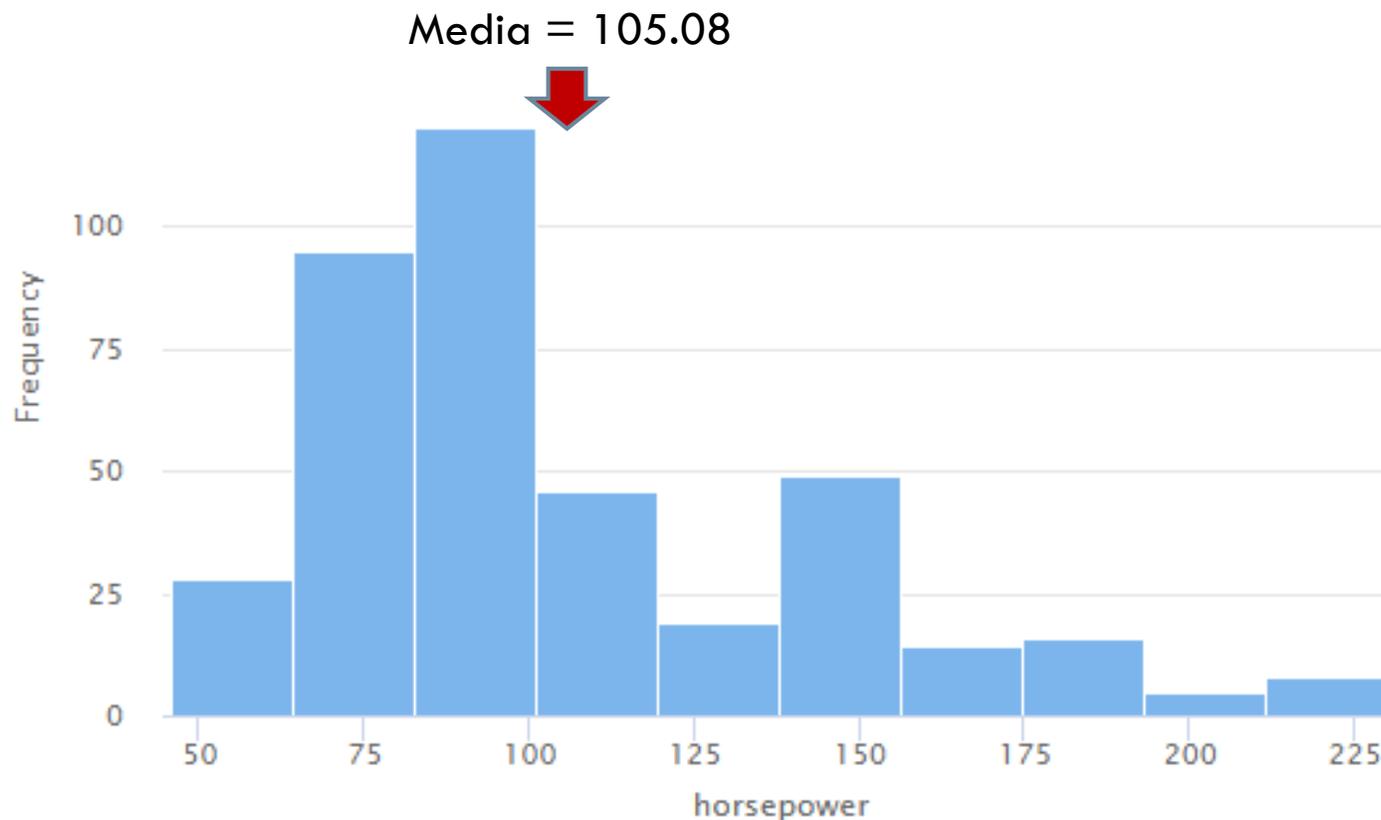
Media 105.08

Mediana 95

Moda 150

Rango medio 138

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



Mediana = 95

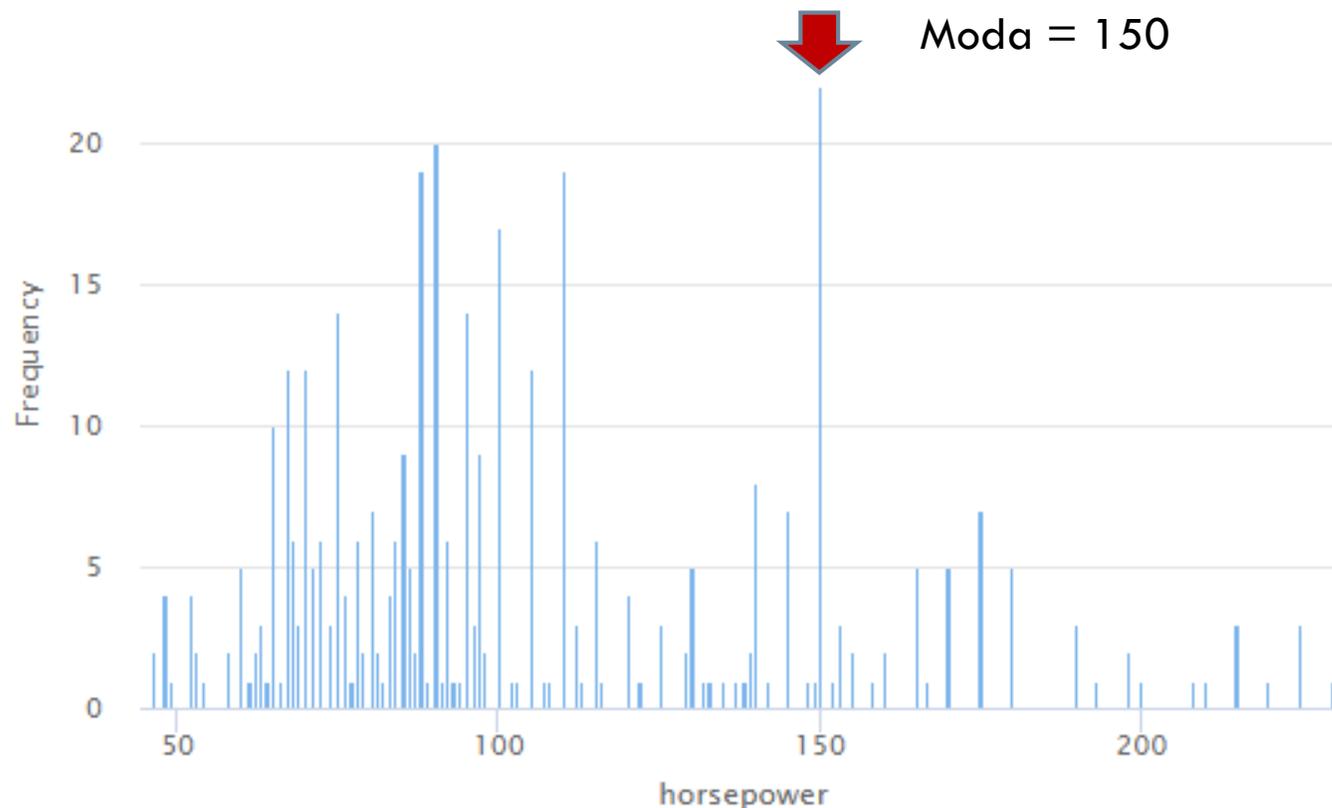
Atributo HORSEPOWER

Media 105.08

Mediana 95

Moda 150

Rango medio 138



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

Mediana = 95

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

MEDIDAS DE DISPERSION

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

VARIANZA Y DESVIACION ESTANDAR

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

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

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

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

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

VARIANZA Y 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 estándar muestral

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

RANGO

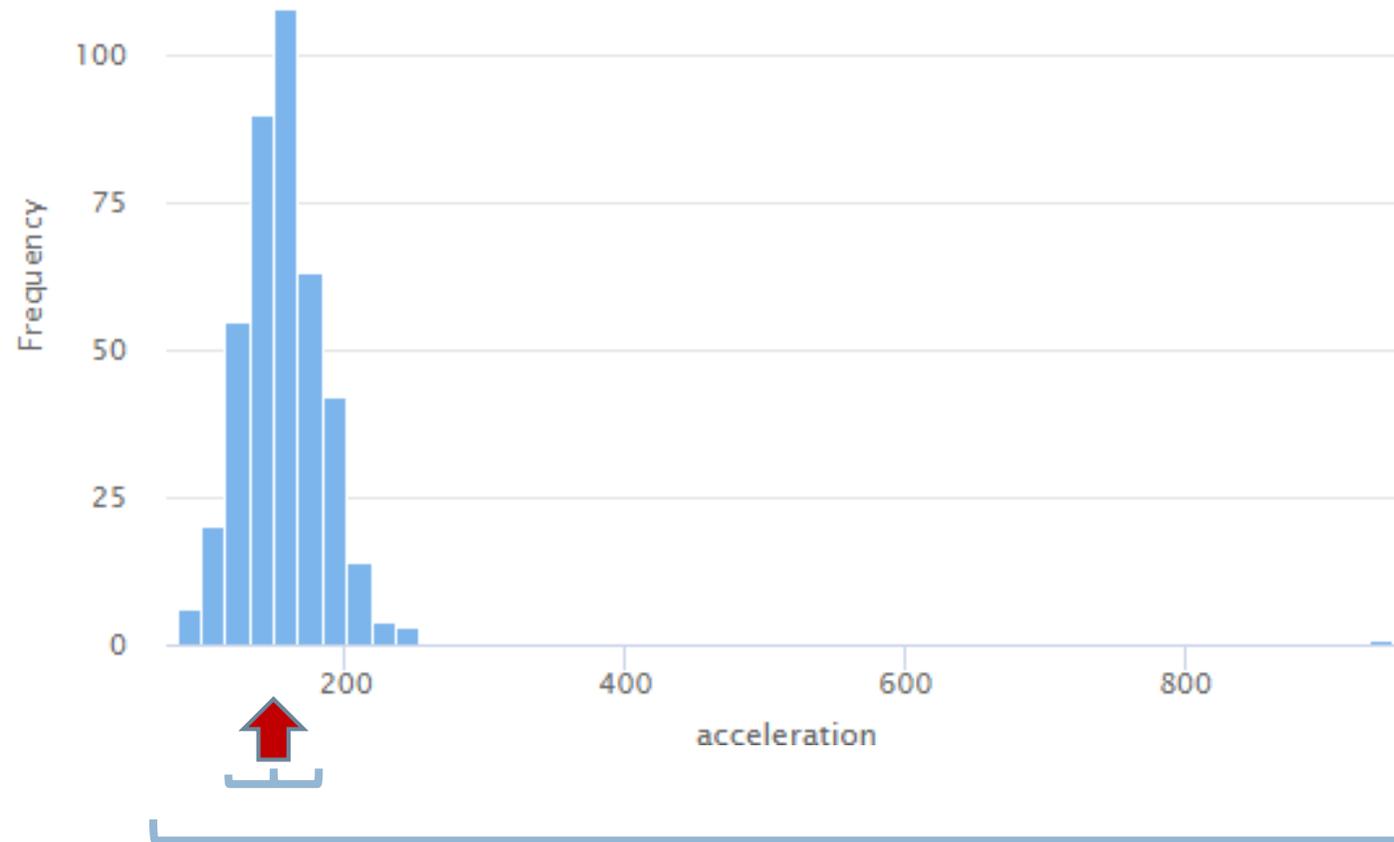
- El rango de un conjunto de valores numéricos es la diferencia entre los valores máximo y mínimo de dicho conjunto.

- Ejemplo

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

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

Atributo ACCELERATION



Media 157.30

Desviación 48.29

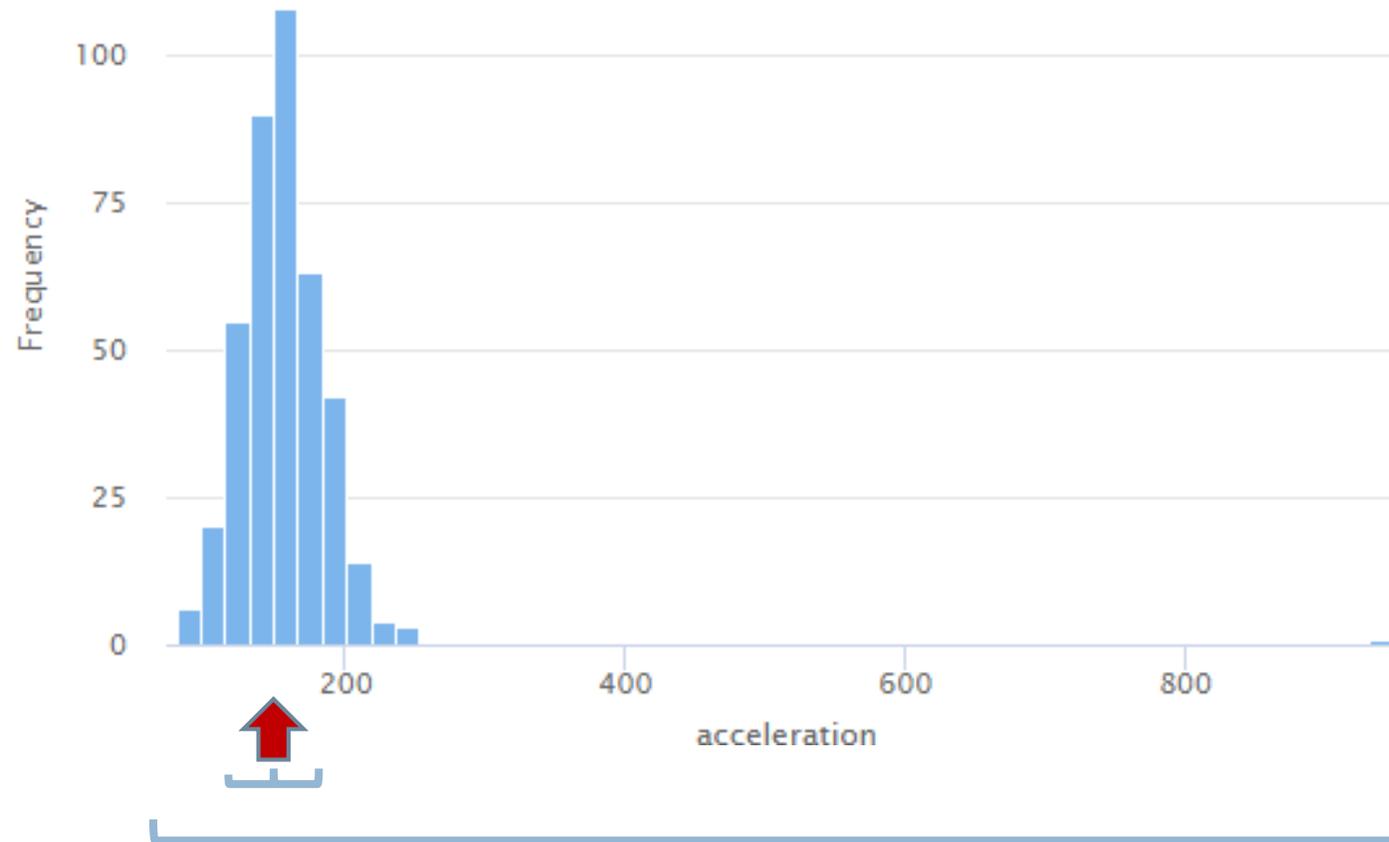
Minimo 80

Maximo 950

Rango 515

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

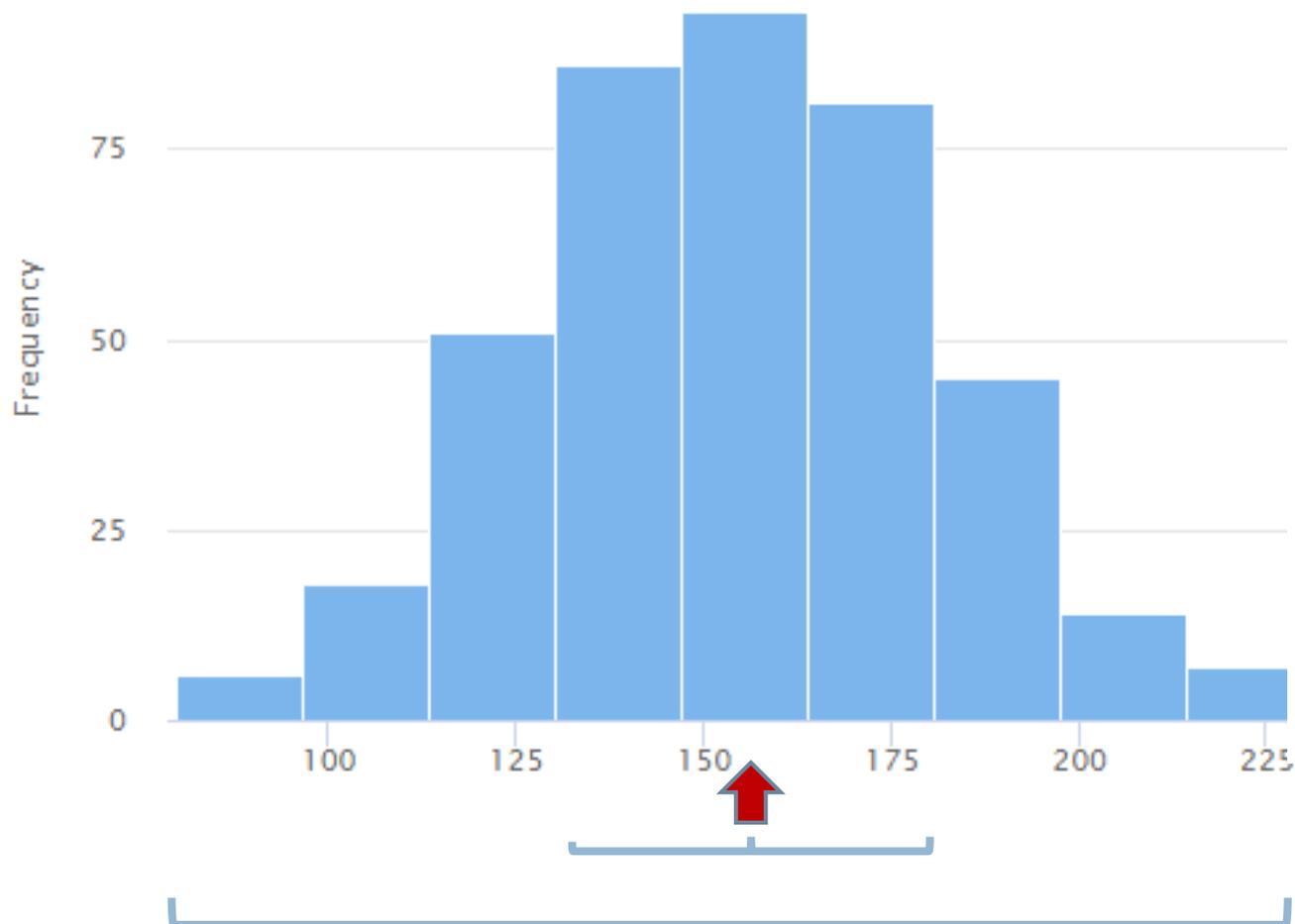
Atributo ACCELERATION



Media	157.30
Desviación	48.29
Minimo	80
Maximo	950
Rango	515

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

Atributo ACCELERATION



Media 155.35

Desviación 27.91

Minimo 80

Maximo 248

Rango 164

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

Cuantiles, Cuartiles y Percentiles

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

□ Ejemplo:

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



$$Q_1 = 47.75$$



$$Q_2 = 54$$



$$Q_3 = 68.25$$

¿Cómo se calculan?

CUARTILES

□ Ejemplo:

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

- La ubicación de Q_1 es $(N+1)/4$ siendo N la cantidad de valores disponibles, 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 Q_3 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

- 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



$$Q_1 = 47.75$$



$$Q_2 = 54$$



$$Q_3 = 68.25$$

RANGO INTERCUARTIL

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

$$RIC = Q_3 - Q_1$$

- Ejemplo:

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

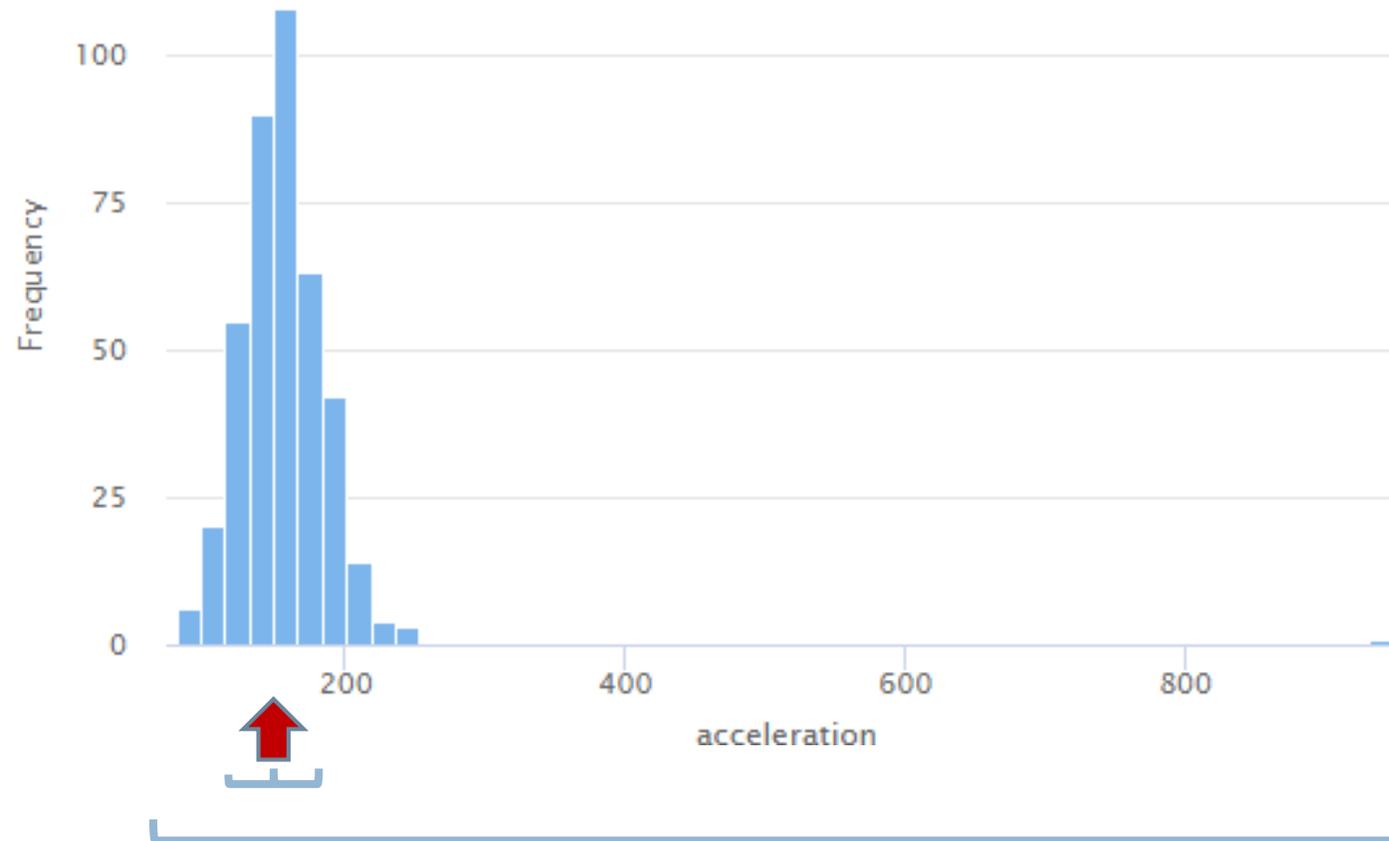

 $Q_1 = 47.75$


 $Q_2 = 54$


 $Q_3 = 68.25$

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

Atributo ACCELERATION



Media	157.30
--------------	---------------

Medidas de Dispersión

Varianza	2332.31
-----------------	----------------

Desviación	48.29
-------------------	--------------

Rango	515
--------------	------------

Q1	137.0
-----------	--------------

Q2	155.0
-----------	--------------

Q3	172.25
-----------	---------------

RIQ	35.25
------------	--------------

Autos-mpg.csv

	mpg	acceleration	horsepower	weight
mean	23.52	157.30	105.08	2979.41
std	7.82	48.29	38.77	847.00
min	9.00	80.00	46.00	1613.00
Q1	17.50	137.00	75.75	2225.25
Q2	23.00	155.00	92.50	2822.50
Q3	29.80	172.25	129.25	3622.50
max	46.60	950.00	230.00	5140.00
RIC	12.30	35.25	53.50	1397.25
Rango	37.60	870.00	184.00	3527.00

Gráfico de Torta/Dona

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: Pie / Donut

Value column: cylinders

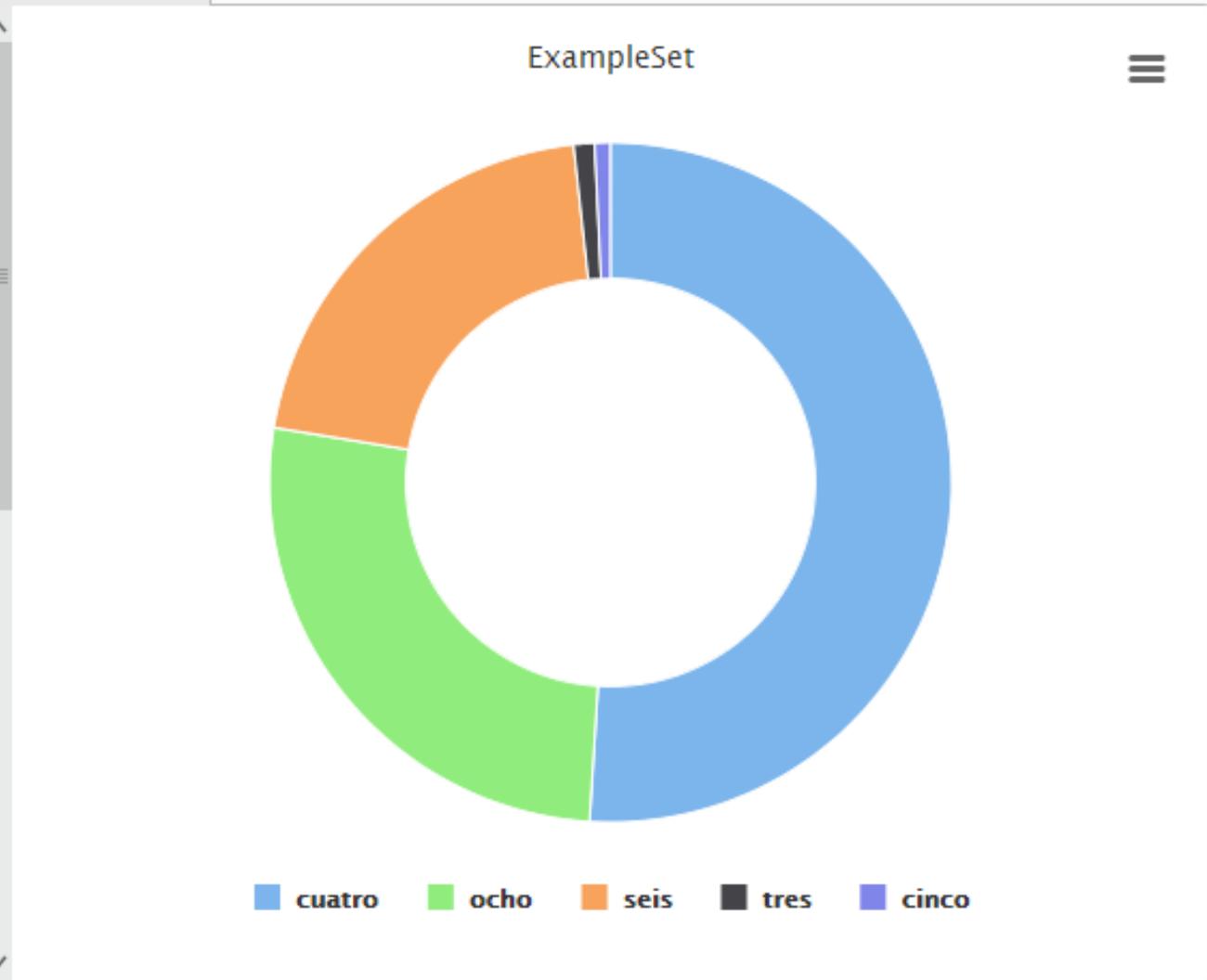
Aggregate data

Group by: cylinders

Aggregation Function: Count

Donut

Plot style >>>



Repository

Import Data

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

Diagrama de dispersión

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: Scatter / Bubble

X-Axis column: horsepower

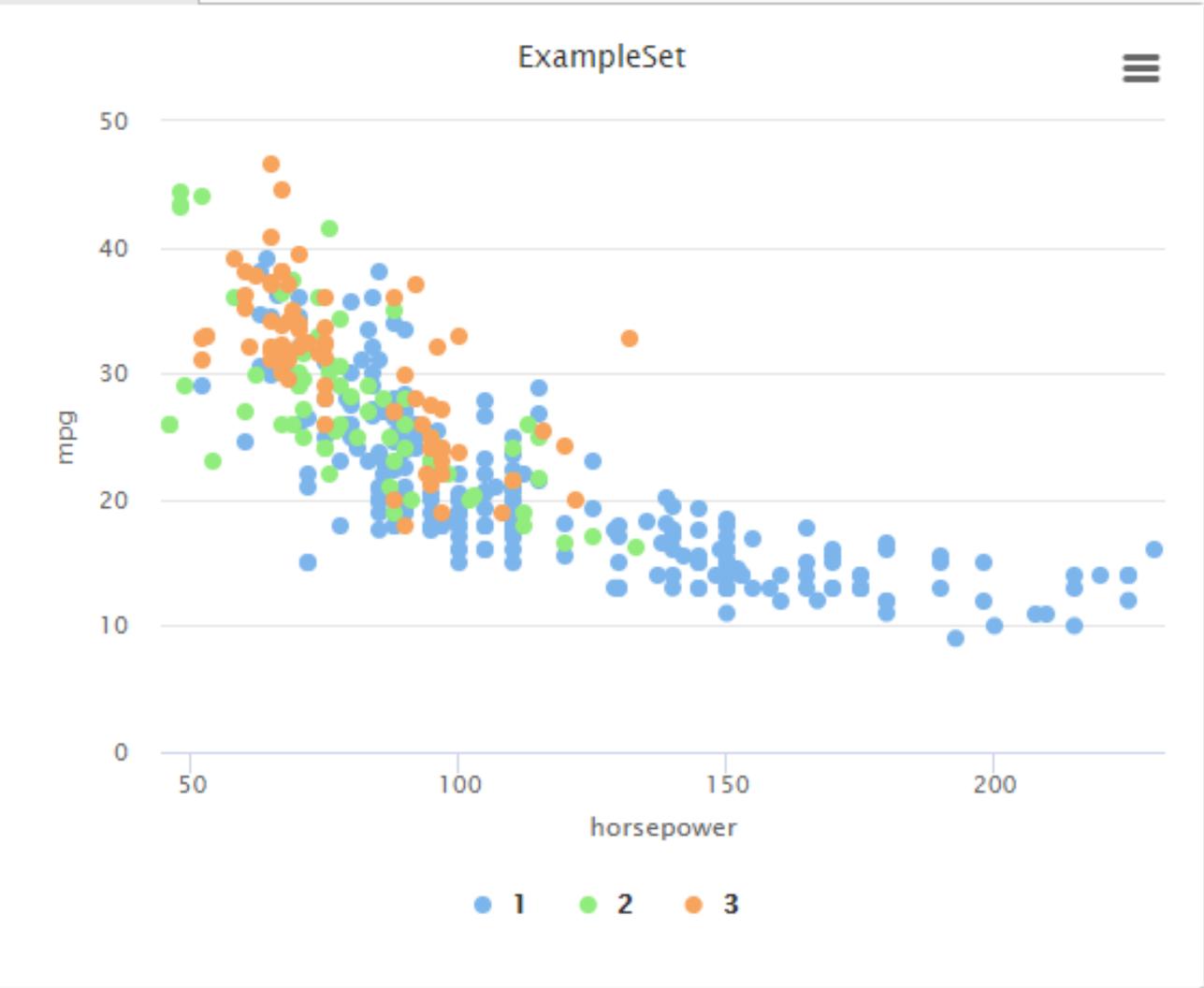
Value column: mpg

Color: origin

Size: -

Jitter: [Slider]

Regression interpolation: None



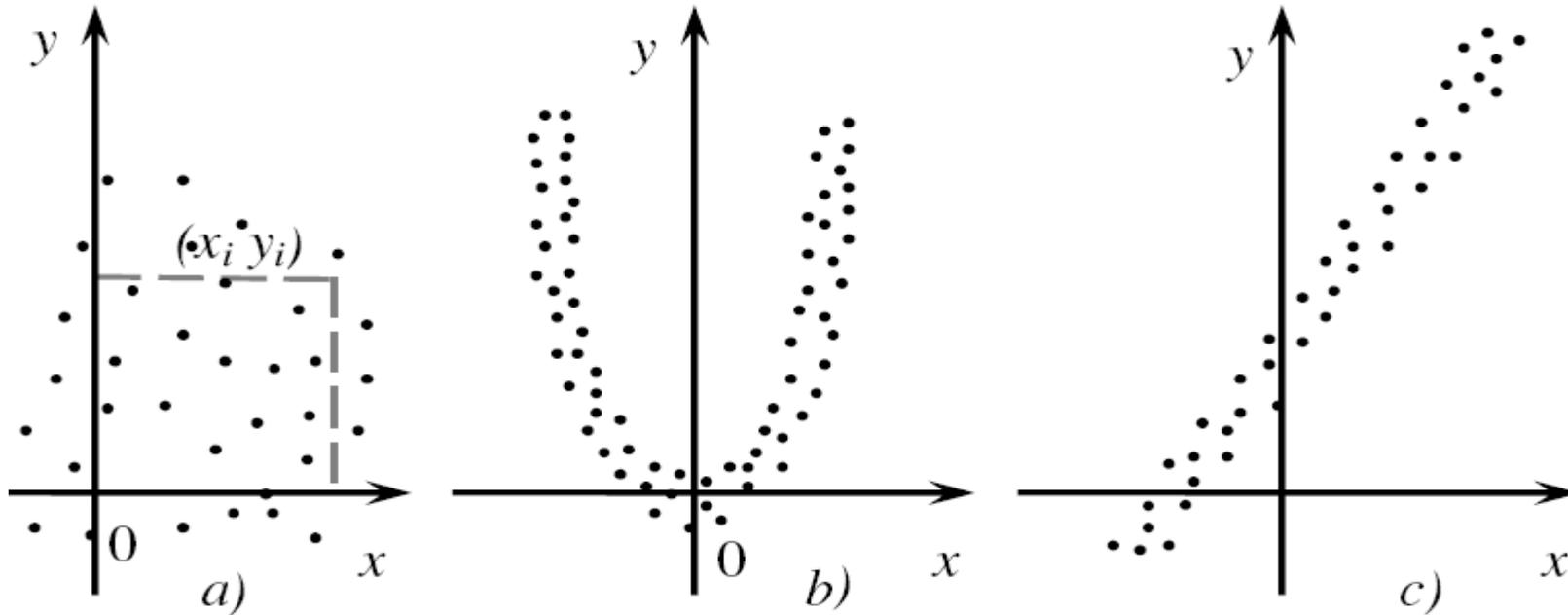
Repository

Import Data

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

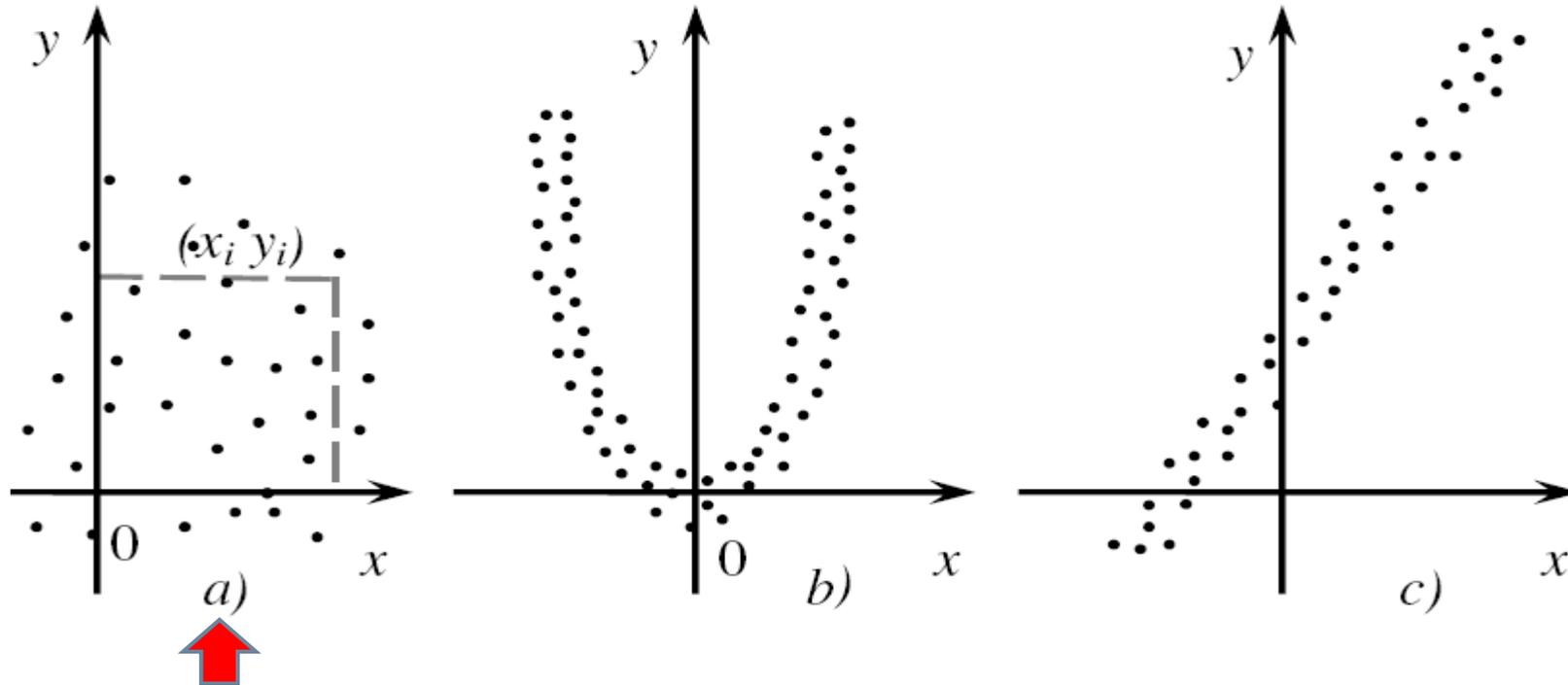
Diagramas de Dispersión

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



Diagramas de Dispersión

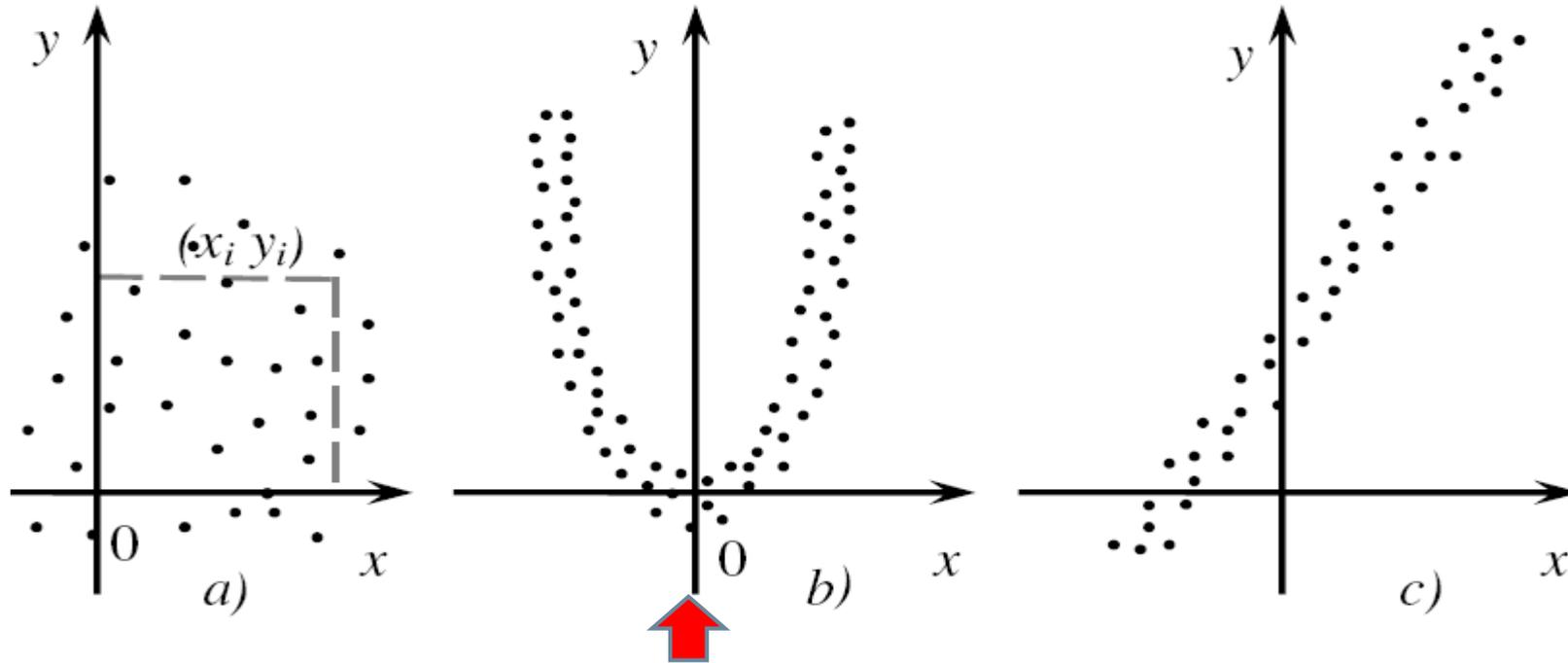
- Consiste en dibujar pares de valores (x_i, y_i) medidos de la v.a. (X, Y) en un sistema de coordenadas



Entre X e Y no hay ninguna relación funcional

Diagramas de Dispersión

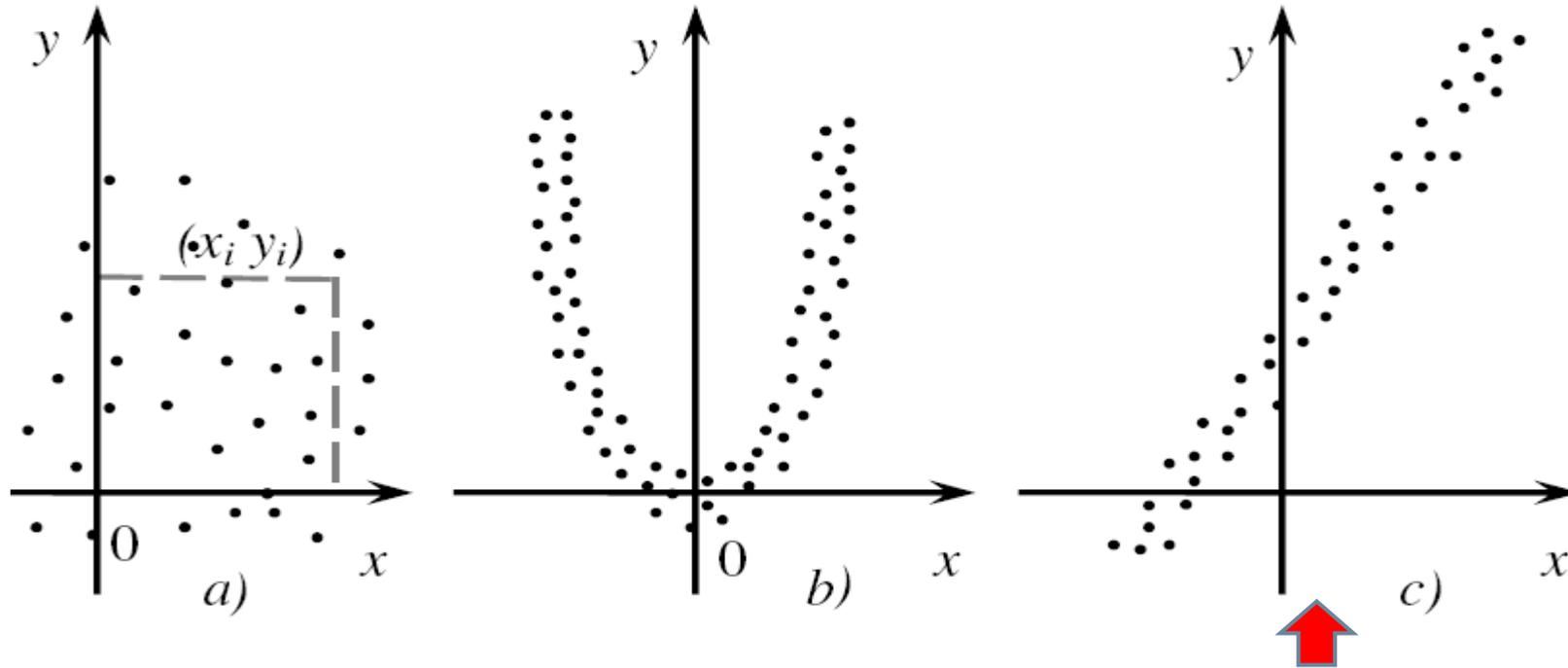
- Consiste en dibujar pares de valores (x_i, y_i) medidos de la v.a. (X, Y) en un sistema de coordenadas



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

Diagramas de Dispersión

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



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

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**.

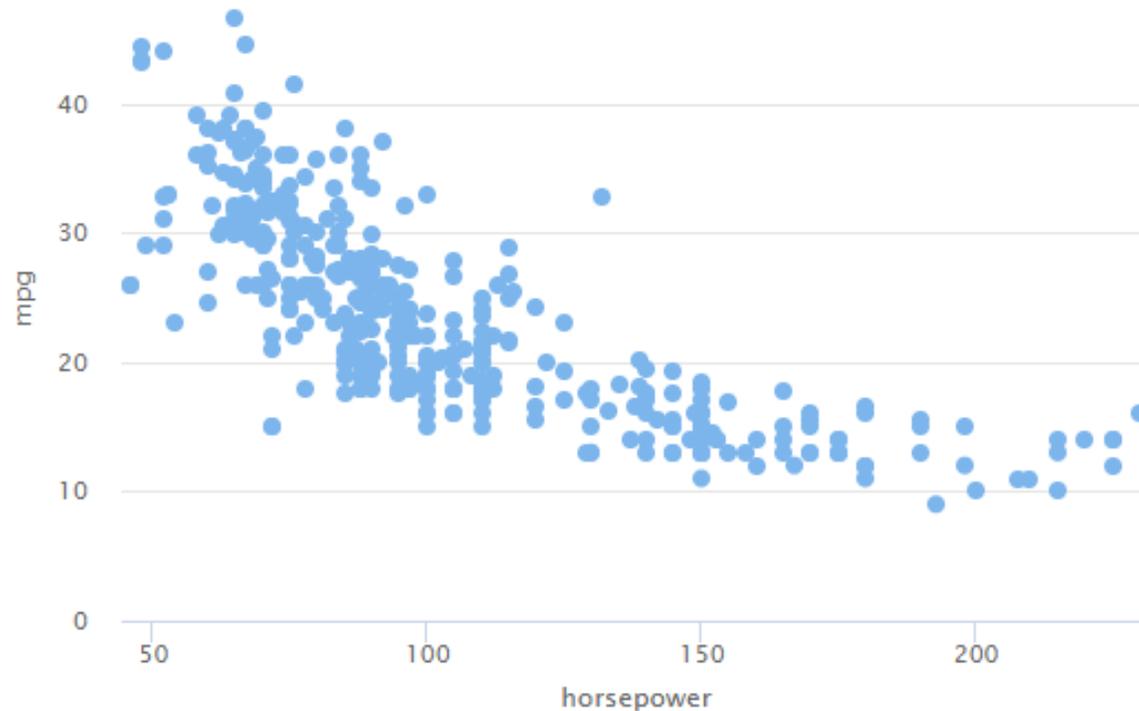


Diagrama de dispersión
entre HORSEPOWER y
MPG

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

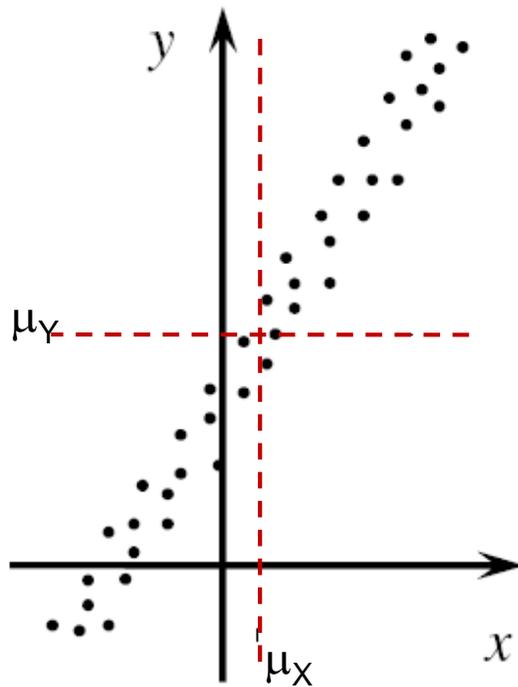
$$\text{Cov}(X, Y) = \left[\sum_{i=1}^N (x_i - \mu_X)(y_i - \mu_Y) \right] / N$$

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

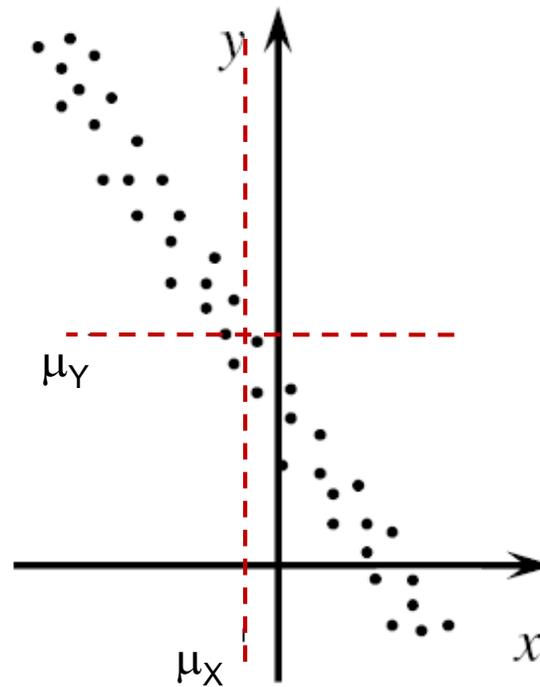
Covarianza

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

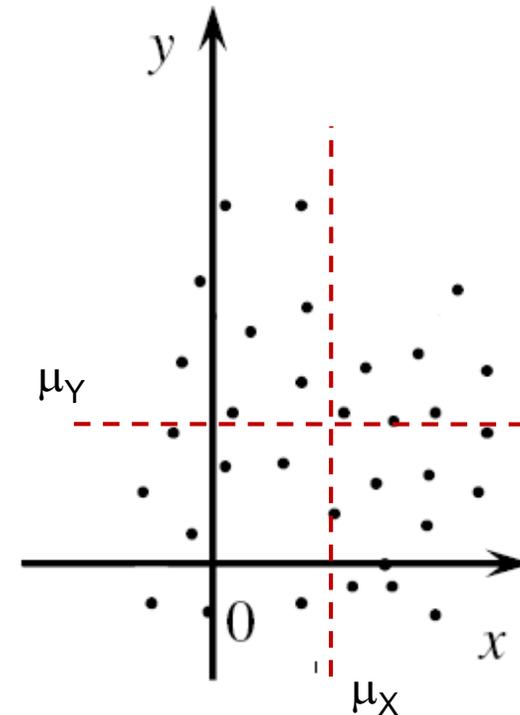
- La **covarianza** es un valor que indica el grado de variación conjunta de dos **variables aleatorias** respecto a sus medias.



Covarianza Positiva



Covarianza Negativa



Covarianza cercana a cero

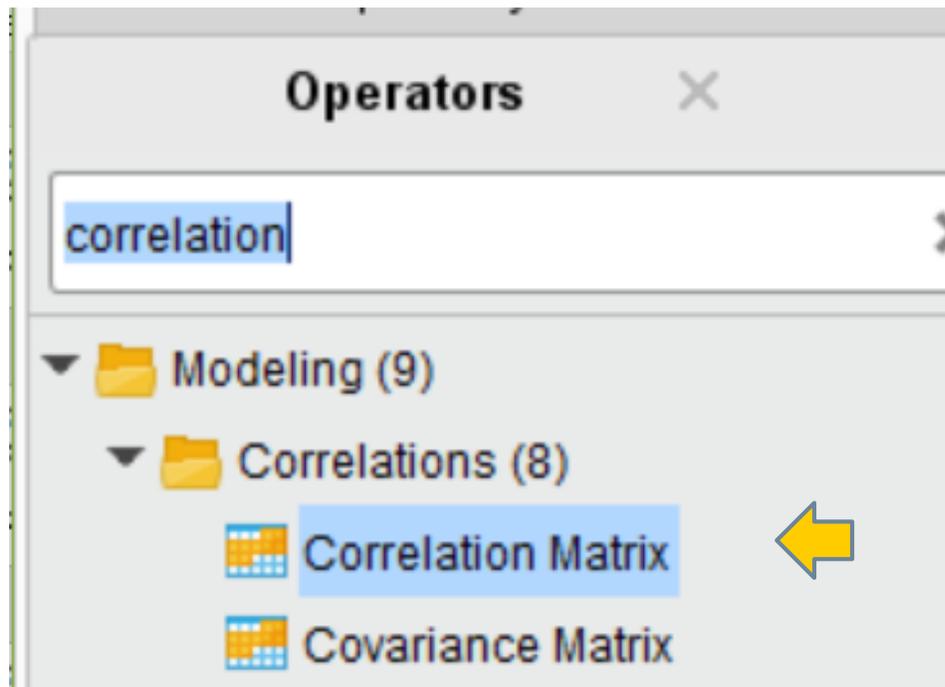
Coeficiente de correlación lineal

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.

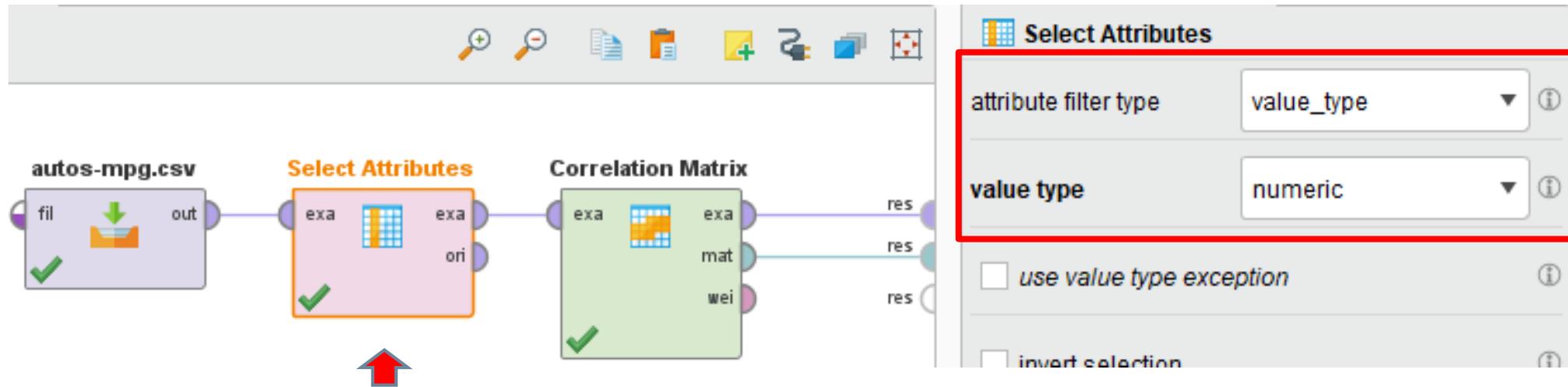
Coeficiente de correlación lineal entre atributos

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



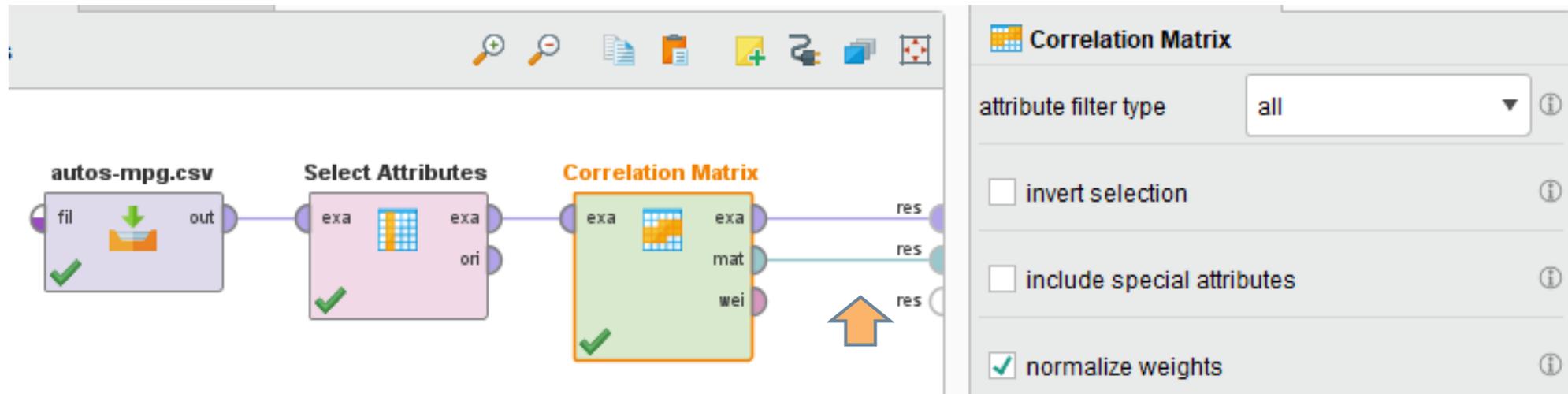
Matriz de Correlación

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

Para obtener esta matriz todos los atributos deben ser numéricos 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
 - ▣ 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.