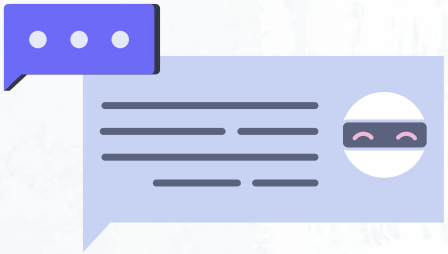
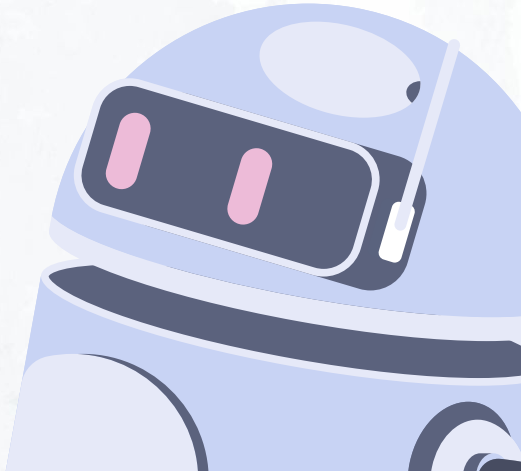
Analítica de Datos (Aprendizaje de máquina)





UNAL

Departamento de Eléctrica, Electrónica y Computación
Facultad de Ingeniería y Arquitectura
Sede Manizales

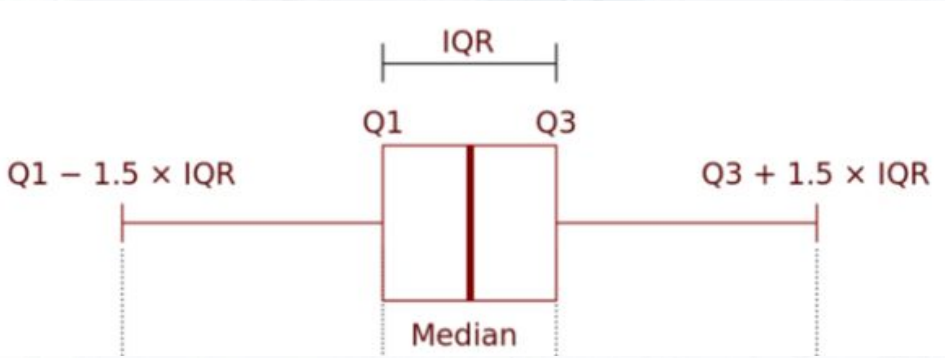
UNIVERSIDAD NACIONAL DE COLOMBIA

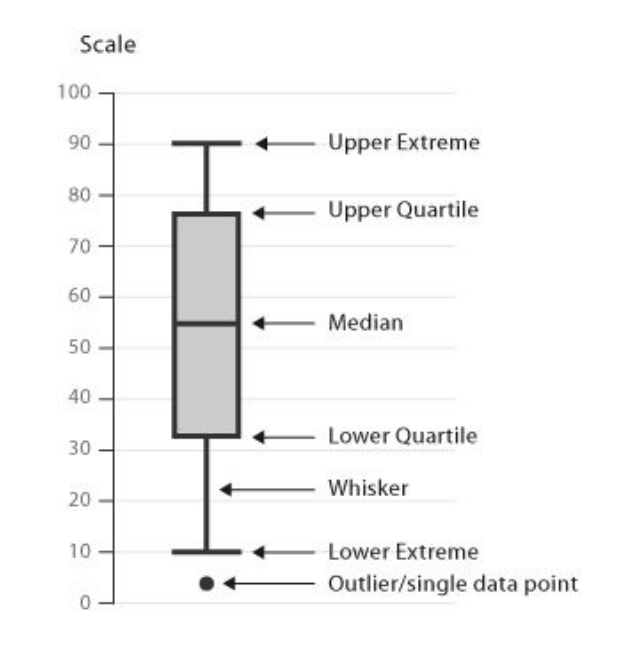
Graficando los datos

Box plots

Una gráfica para resumirlo todo.

También conocida como Box and Whisker Plot (gráfico de caja y bigotes)



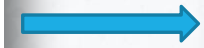


Graficando los datos

Box plots

Ordenamos los valores de menor a mayor

```
persona;valor_casa
carolina;10
ariana;12
guillermo;13
christian;9
agustina;14
sebastian;12
gabriel;11
lucia;10
gustavo;9.5
damiano;8.3
elisa;10.6
carlos;12
german;13
tomas;7.5
braian;9
maria;10.2
micaela;11.7
marcelo;12.5
```



```
7.5
8.3
9
9
9.5
10
10
10.2
10.6
11
11.7
12
12
12
12.5
13
13
14
```

$$\frac{k \cdot (n+1)}{4} \qquad k = 1, 2, 3$$

Graficando los datos

Box plots (incluyamos ahora un billonario en el análisis)

```
persona;valor_casa
carolina;10
ariana;12
guillermo;13
christian;9
agustina;14
sebastian;12
gabriel;11
lucia;10
gustavo;9.5
damiano;8.3
elisa;10.6
carlos;12
german;13
tomas;7.5
braian;9
maria;10.2
micaela;11.7
marcelo;12.5
elon;800
```

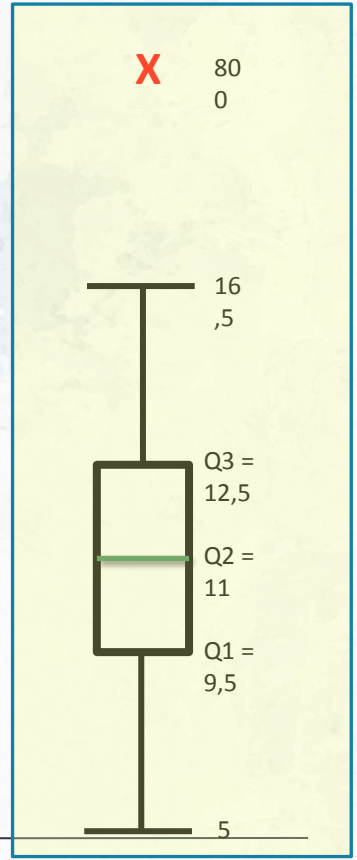
```
7.5
8.3
9
9
9.5
10
10
10.2
10.6
11
11.7
12
12
12
12.5
13
13
14
800
```

media => *mean({x})* = 52,384
desviación => *std({x})* = 176,223
mediana => *median({x})* = 11 = Q2

Q1 => 9,5
Q3 => 12,5

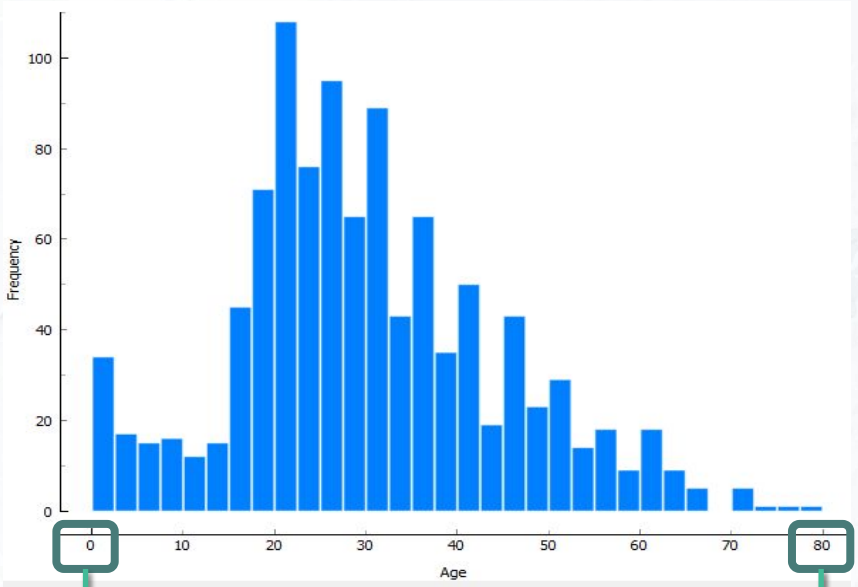
Rango intercuartil=> *iqr({x})* = 12,5 – 9,5 = 3

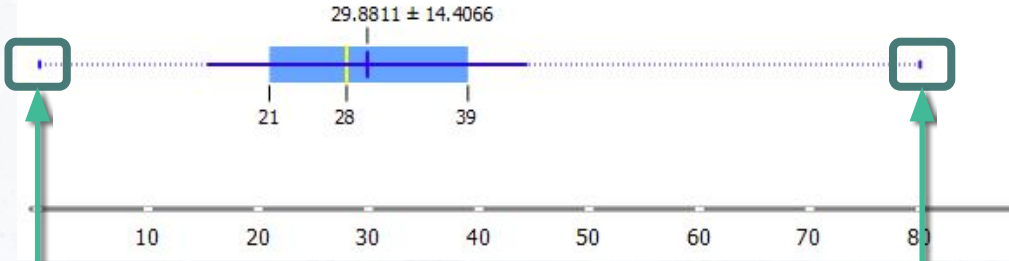
Bigote inferior => Q1 – (1,5 * iqr) = 5
Bigote superior=> Q3 + (1,5 * iqr) = 16,5



Graficando los datos

Box plots Ejemplos en Orange Data Mining

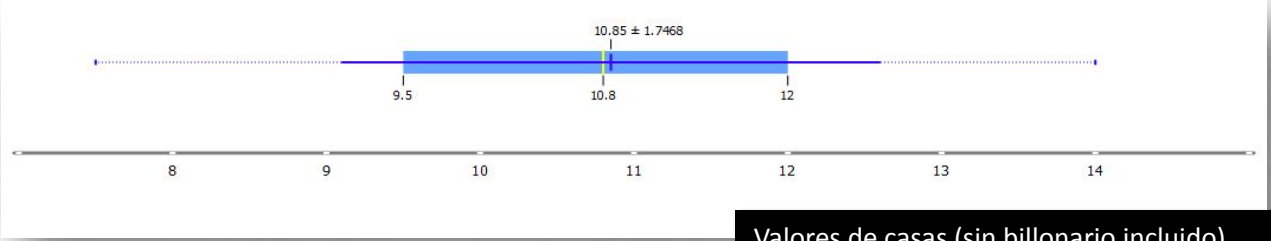




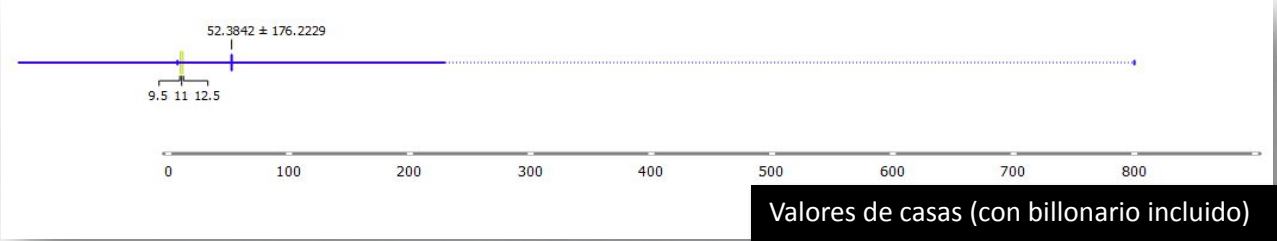
Valor mínimo:	0.16
Valor máximo:	80
Media:	29.88
Desviación estándar:	14.40
Mediana:	28

Graficando los datos

Box plots Ejemplos en Orange Data Mining



Valores de casas (sin billonario incluido)

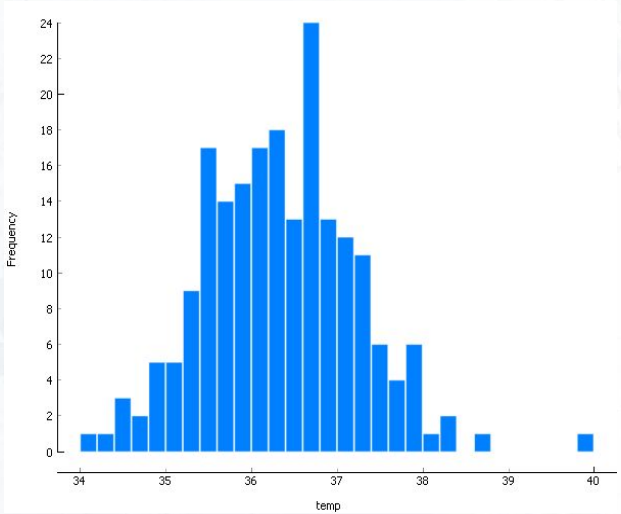


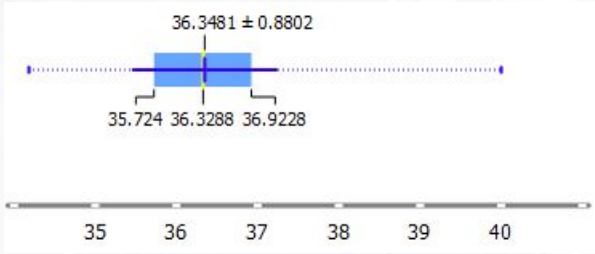
Valores de casas (con billonario incluido)

Graficando los datos

¿Qué nos dicen los **histogramas**?

- Dan cuenta de la distribución de los datos.
- Los picos indican la moda(s).
- Permiten comprobar si las colas (en los extremos) tienen datos poco comunes.
- **No siempre simétricos**

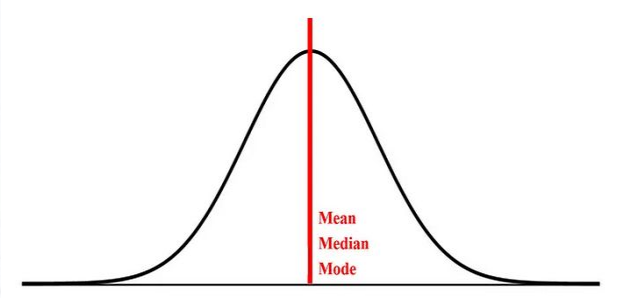




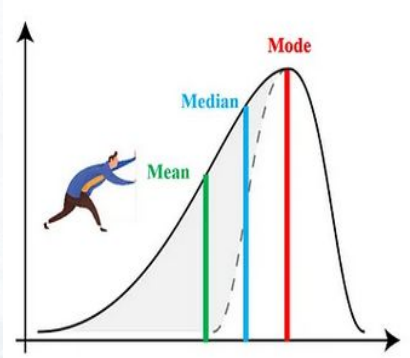
UNAL
Departamento de Eléctrica, Electrónica y Computación
Facultad de Ingeniería y Arquitectura
Sede Manizales
UNIVERSIDAD NACIONAL DE COLOMBIA

Graficando los datos

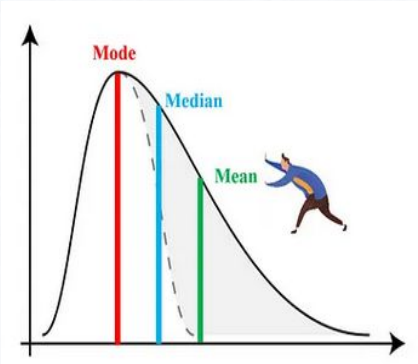
¿Qué nos dicen los **histogramas**?
Asimetría (skewness)



Distribución
Normal: **Simétrica**



Negative Skew

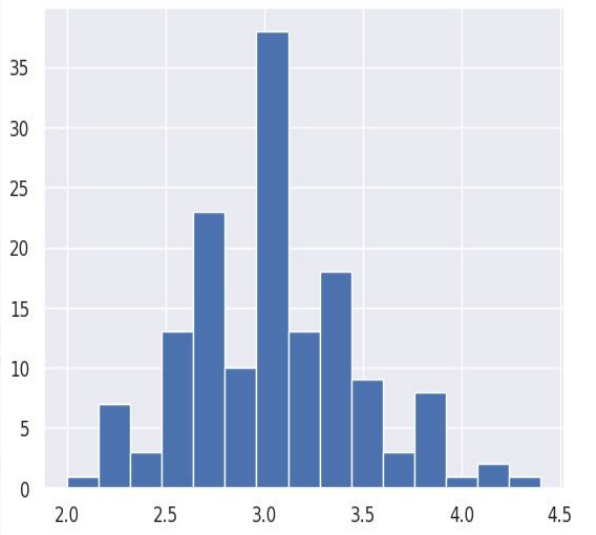


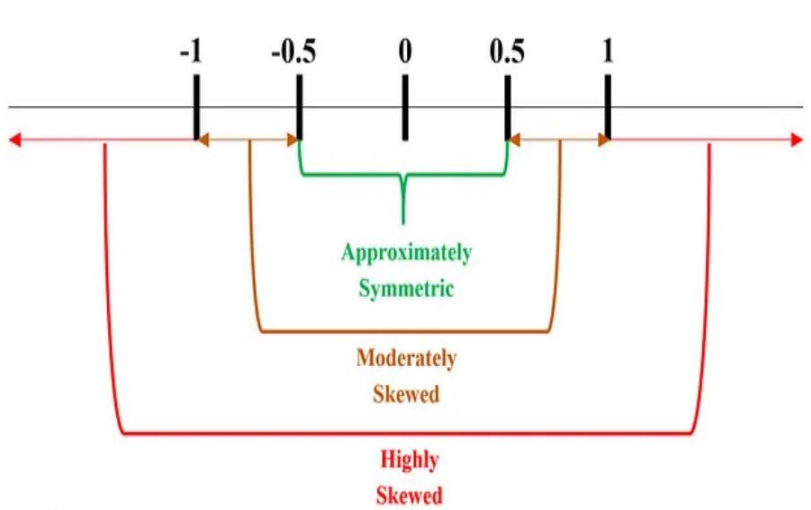
Positive Skew

UNAL
Departamento de Eléctrica, Electrónica y Computación
Facultad de Ingeniería y Arquitectura
Sede Manizales
UNIVERSIDAD NACIONAL DE COLOMBIA

Graficando los datos

¿Qué nos dicen los **histogramas**?
Asimetría (skewness)

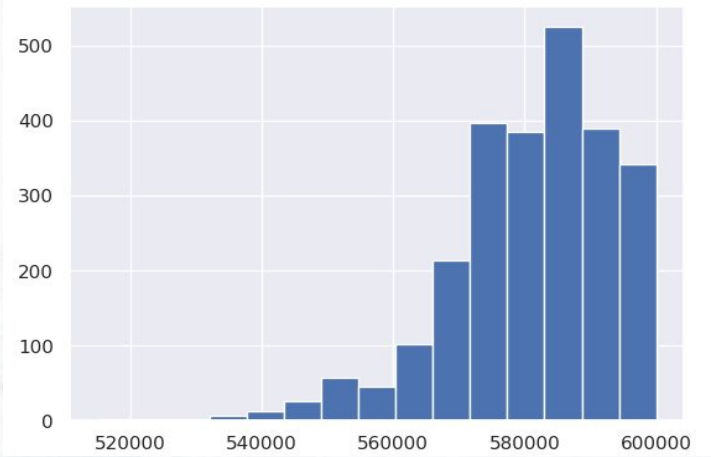




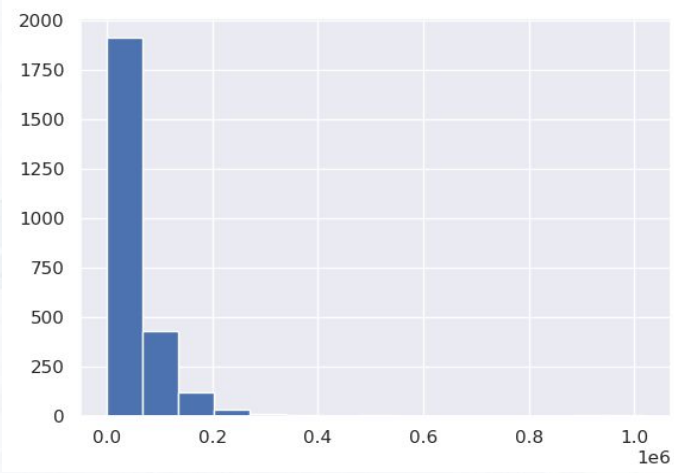
Skew: 0.33 => Aproximadamente simétrica

Graficando los datos

¿Qué nos dicen los **histogramas**?
Asimetría (skewness)



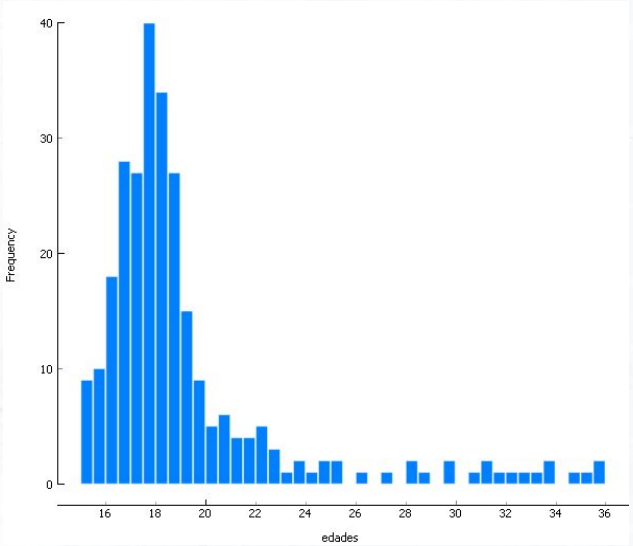
Skew: -0.922



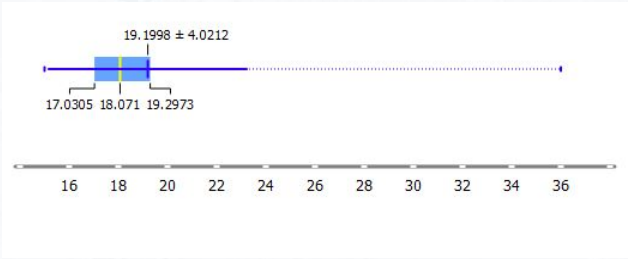
Skew: 7.075

Graficando los datos

¿Qué nos dicen los **histogramas**?
Asimetría (skewness)



El sesgo se hace evidente en los histogramas y en los box-plots

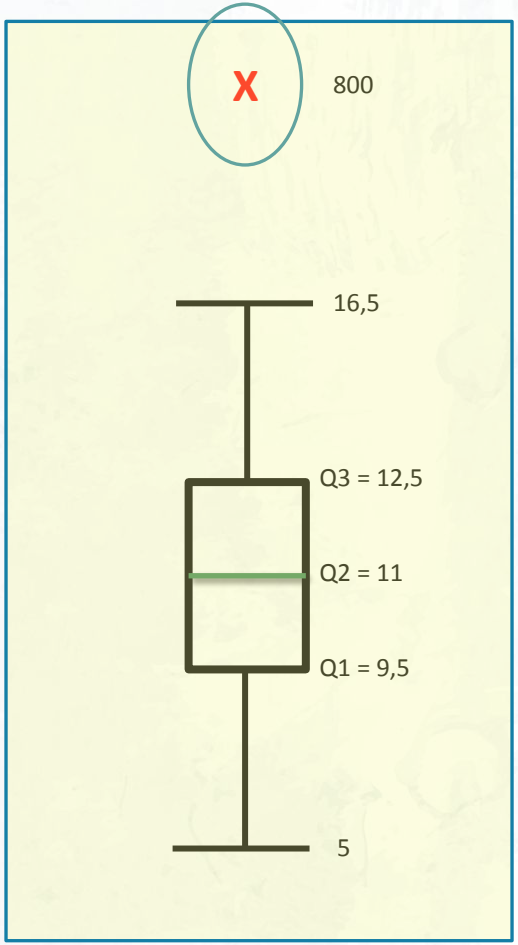


Graficando los datos

Valores atípicos

Son valores de la muestra que se escapan de los valores esperados

- Muy grandes o muy pequeños
- Baja frecuencia
- Pueden resultar muy dañinos en los cálculos estadísticos

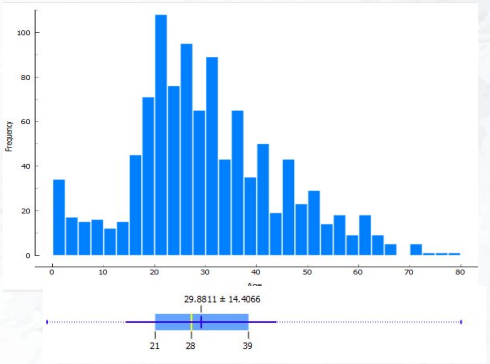


Graficando los datos

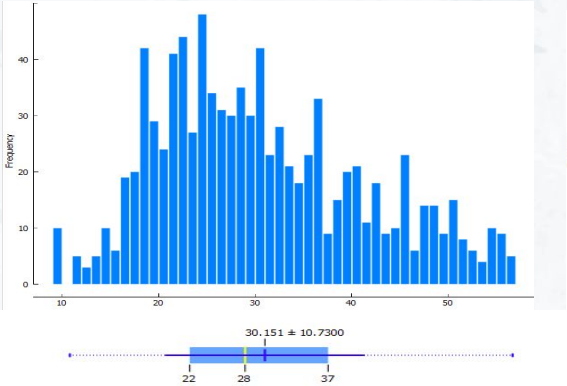
Valores atípicos - Detección

- Visualmente: cola en los histogramas, Boxplots (1.5*IQR)
- Límite fijo: Media ± 3σ
- Otros métodos de clasificación

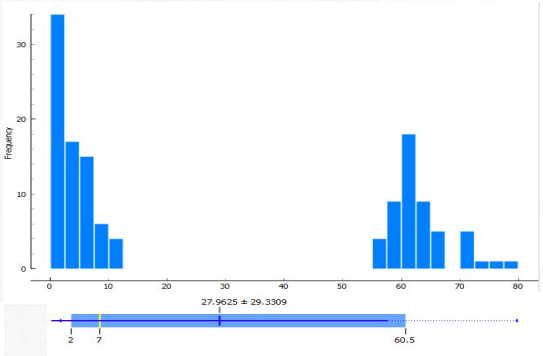
Distribución completa



Inliers

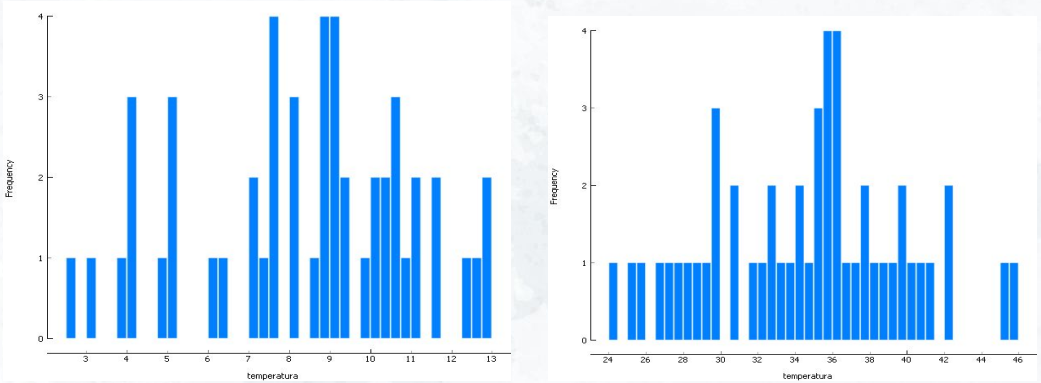


Outliers



Graficando los datos

Normalización de los datos



Temperatura en invierno vs. Temperatura en verano

Los **histogramas** nos pueden aportar información útil de nuestras variables, sin embargo, pueden ser difíciles de comparar si estamos analizando variables en distintas unidades o tomadas bajos condiciones distintas.

Ejemplos de la necesidad de normalización de datos:

El histograma de longitudes vendrá en metros. El histograma de masa vendrá en kilogramos.

UNAL
Departamento de Eléctrica, Electrónica y Computación
Facultad de Ingeniería y Arquitectura
Sede Manizales
UNIVERSIDAD NACIONAL DE COLOMBIA

Graficando los datos

Normalización de los datos

Coordenadas estándar o valores estándar o valores normalizados (o simplemente z-scores)

Asumamos que tenemos un conjunto de datos ***{x}*** compuesto por ***N*** elementos, x_1, x_2, …, x_N.

Los **valores normalizados** de ese conjunto *{x}* es igual a:

$$\hat{x}_i = \frac{(x_i - \text{mean}(\{x\}))}{\text{std}(\{x\})}.$$

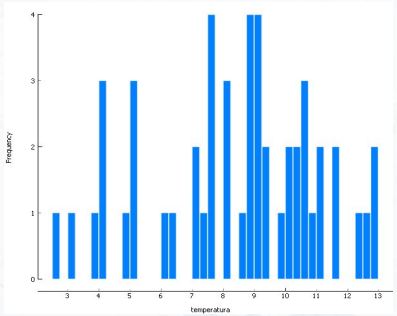
Con esto conseguiremos que

$$\text{mean}(\{\hat{x}\}) = 0. \qquad \text{std}(\{\hat{x}\}) = 1.$$

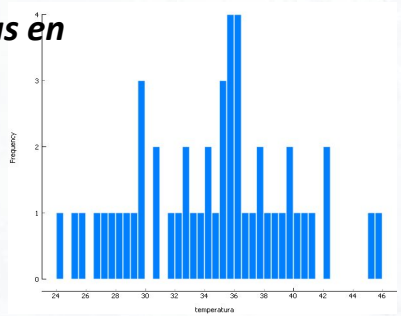
Graficando los datos

Normalización de los datos

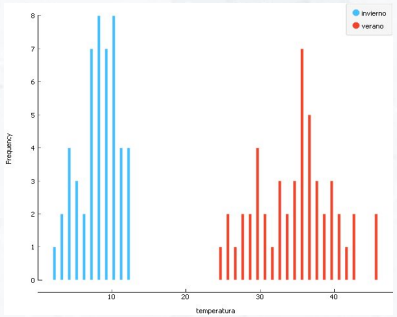
Temperaturas en invierno



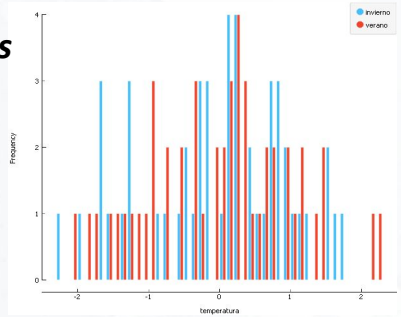
Temperaturas en verano



Sin normalizar



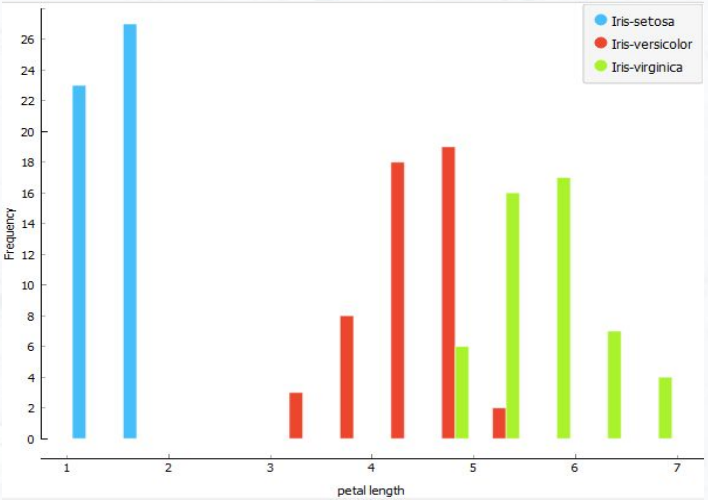
Valores normalizados



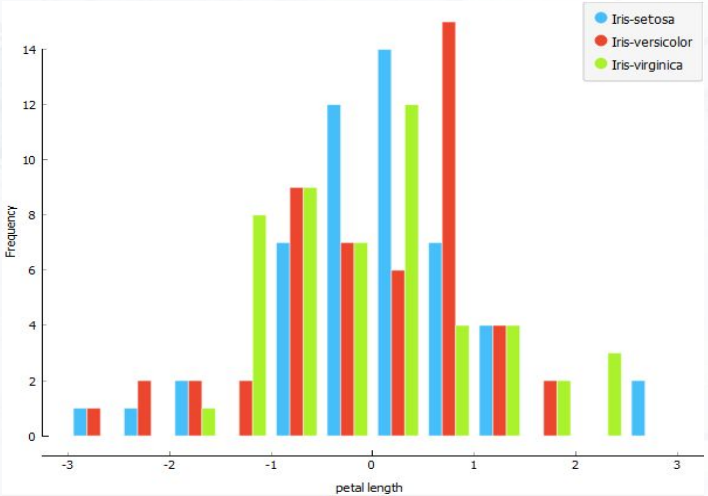
Graficando los datos

Normalización de los datos

Sin normalizar. Largo del pétalo



Valores normalizados (z-scores)



Graficando datos

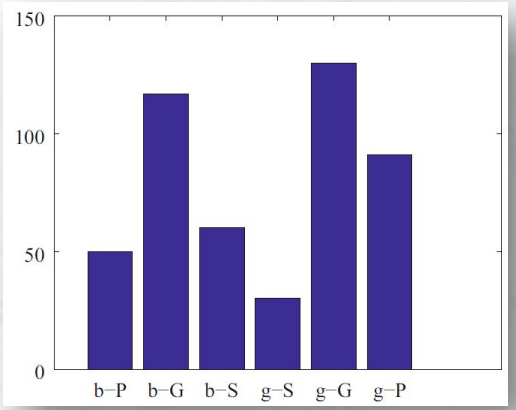
Buscando relaciones

- Hasta ahora hemos analizado una única variable en nuestros datos.
- Sin embargo, uno de los fines del análisis y modelamiento de datos es encontrar relaciones entre las distintas variables que conforman los datasets bajo estudio y partir de estas relaciones poder inferir dependencias, causalidad e impactos.
- Veamos algunas de las técnicas de visualización y análisis descriptivo de dos variables que se emplean a la hora de describir y explorar los datos.

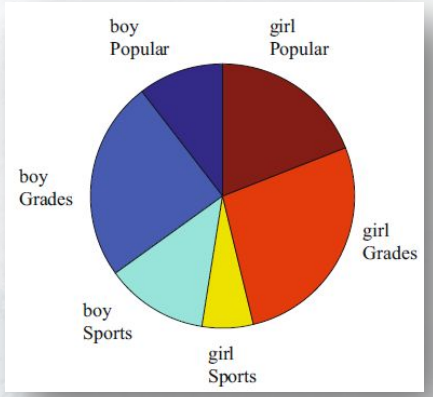
Graficando datos

Buscando relaciones

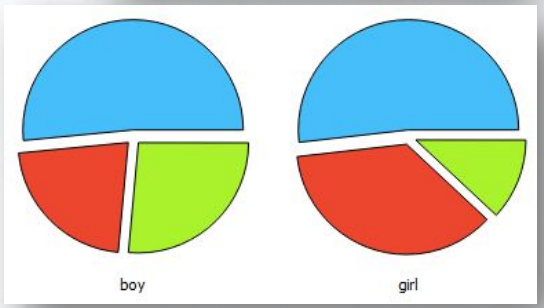
- Datos categóricos (conteos y cuadros)



Gráfica de barras



Gráfica de Torta



“género"	“objetivo"
niño	Deportes
niño	Popular
niña	Popular
niña	Popular
niña	Popular
niña	Popular
niña	Popular
niña	Notas
niña	Deportes
niña	Deportes
niña	Deportes
niña	Notas
niño	Popular
niño	Popular
niño	Popular
niña	Notas
niña	Deportes
niña	Popular
niña	Notas
niña	Popular
niña	Popular
niña	Notas
niña	Popular
niño	Notas
niño	Notas
niño	Notas
niño	Notas
niño	Notas
niño	Deportes
...	...

UNAL

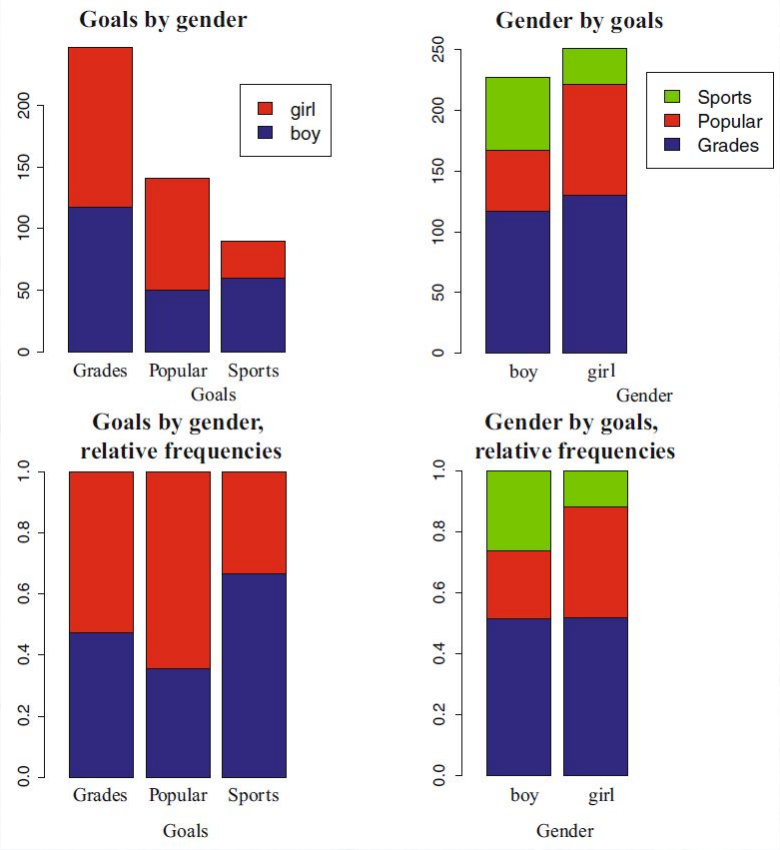
Departamento de Eléctrica, Electrónica y Computación
Facultad de Ingeniería y Arquitectura
Sede Manizales

UNIVERSIDAD NACIONAL DE COLOMBIA

Graficando datos

Buscando relaciones

Stacked bars (barras apiladas)



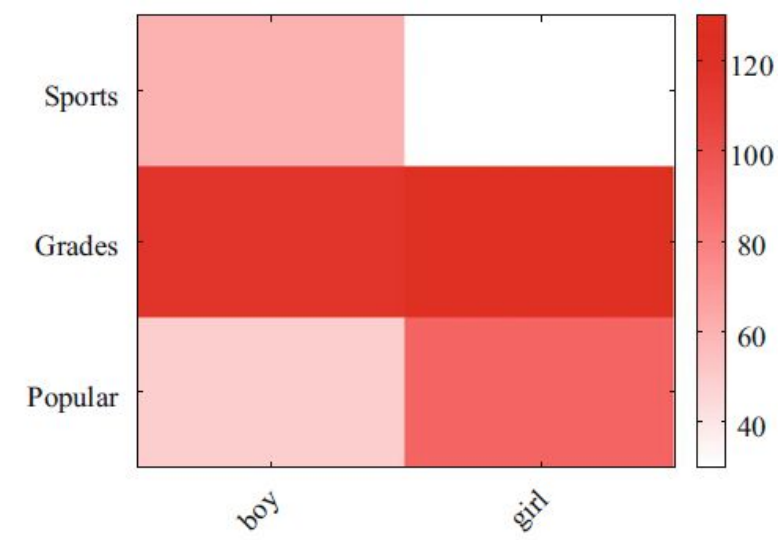
Departamento de Eléctrica, Electrónica y Computación
Facultad de Ingeniería y Arquitectura
Sede Manizales

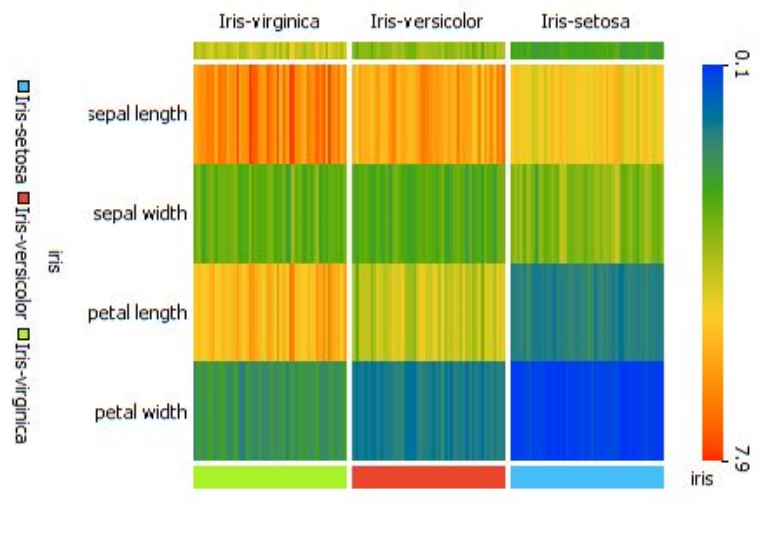
UNIVERSIDAD NACIONAL DE COLOMBIA

Graficando datos

Buscando relaciones

Mapas de calor (Heatmaps)

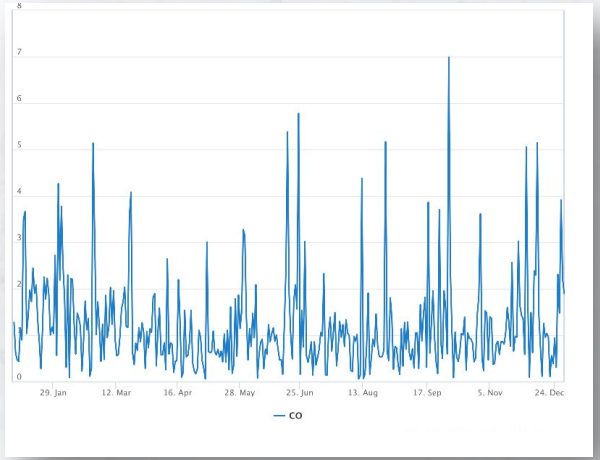


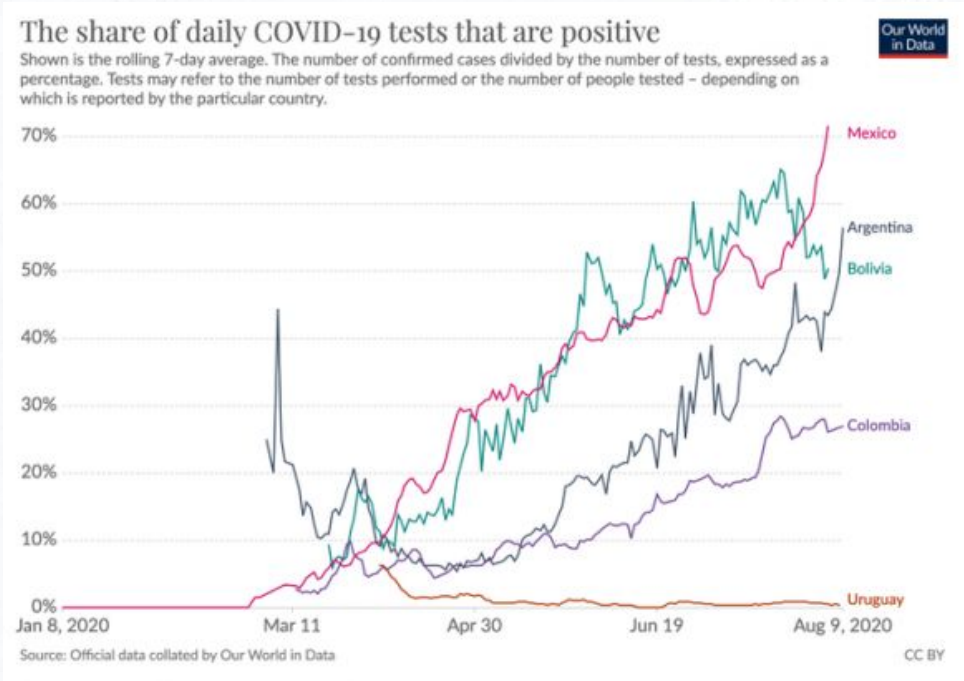


Graficando datos

Buscando relaciones

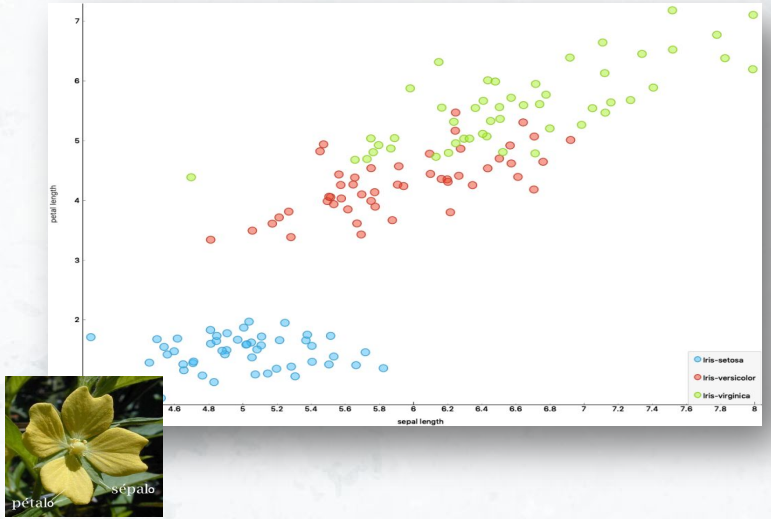
- Series temporales





Graficando datos

Buscando relaciones



Gráficos de dispersión
(Scatter plots)

En los gráficos de dispersión se grafica la relación entre dos atributos numéricos.

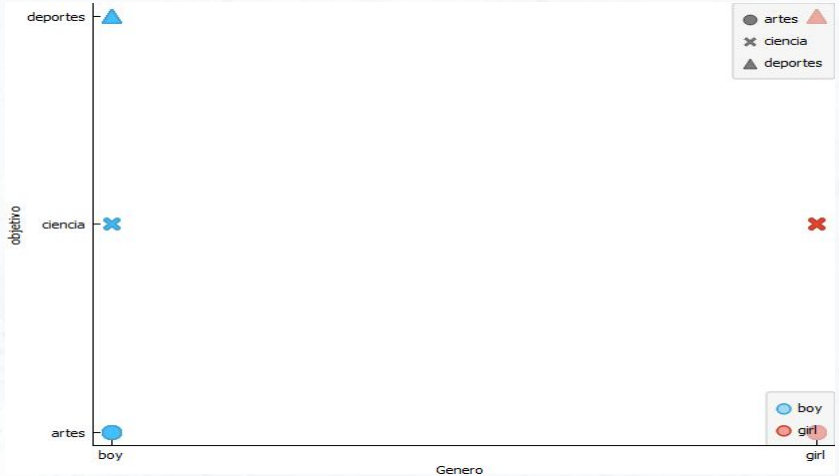
De un scatter plot se puede inferir el tipo **(lineal o no lineal)** y la intensidad de relación entre variables.

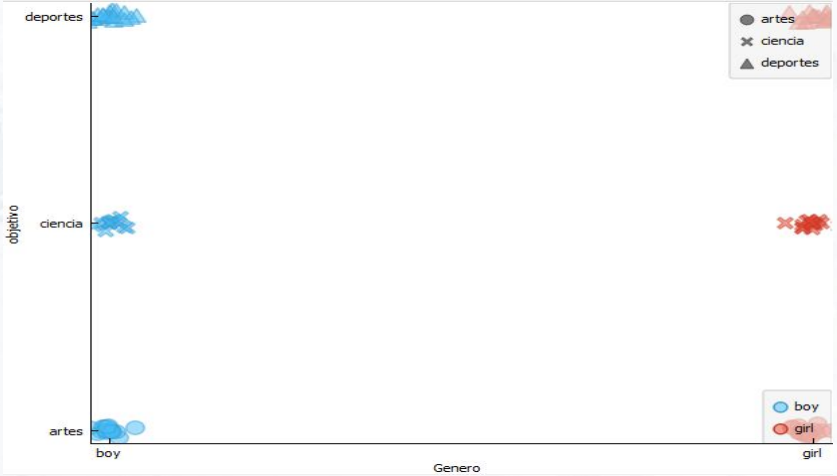
Se puede agregar más información a un scatter plot en forma de una tercera dimensión o en forma de etiquetas de colores.

Graficando datos

Buscando relaciones

Gráficos de dispersión (Scatter plots) - **Variables categóricas**



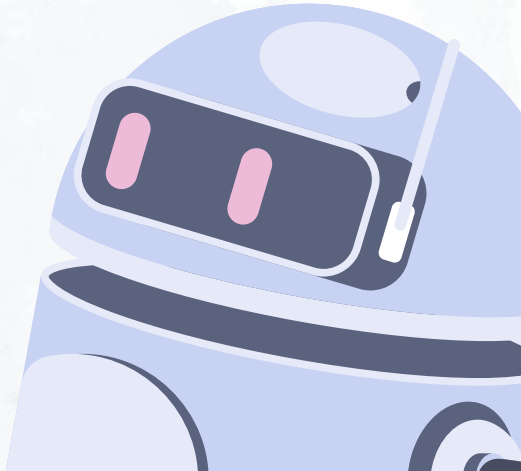


Tarea - Consulta

- ¿Qué es la correlación?
- Causalidad vs. Correlación
- Mapas de Correlación
- Correlación para visualización

Gracias !

dfcollazosh@unal.edu.co



UNAL

Departamento de Eléctrica, Electrónica y Computación
Facultad de Ingeniería y Arquitectura
Sede Manizales

UNIVERSIDAD
NACIONAL
DE COLOMBIA