Introducción
Análisis exploratorio de datos

Introducción a la Bioingeniería

Grado en Ingeniería en Sistemas de Telecomunicación

Luis Bote Curiel

Contenido

1. Introducción
2. Herramientas para EDA
3. Tipos de variables
4. Tipos de EDA
5. Conjunto de datos

Introducción

- El **análisis exploratorio de datos** (*exploratory data analisis*, EDA) se utiliza en la ciencia de datos (*data science*) para **analizar e investigar conjuntos de datos** y **resumir sus principales características**, a menudo empleando métodos de visualización de datos.
- El **objetivo** principal del EDA es **analizar los datos** antes de hacer suposiciones.
- **Puede ayudar a entender los datos, identificar errores**, **detectar valores atípicos** o anómalos y encontrar **relaciones** interesantes **entre las variables**.

Introducción

- **Una vez que se ha completado** el EDA y se han extraído conclusiones, sus resultados pueden utilizarse para **análisis o modelado de datos más sofisticados**, incluido el **aprendizaje automático**.

Herramientas para EDA

- **Python**
  - **Pandas**
  - **Numpy**
  - **Matplotlib**
  - SciPy
- R
- Matlab

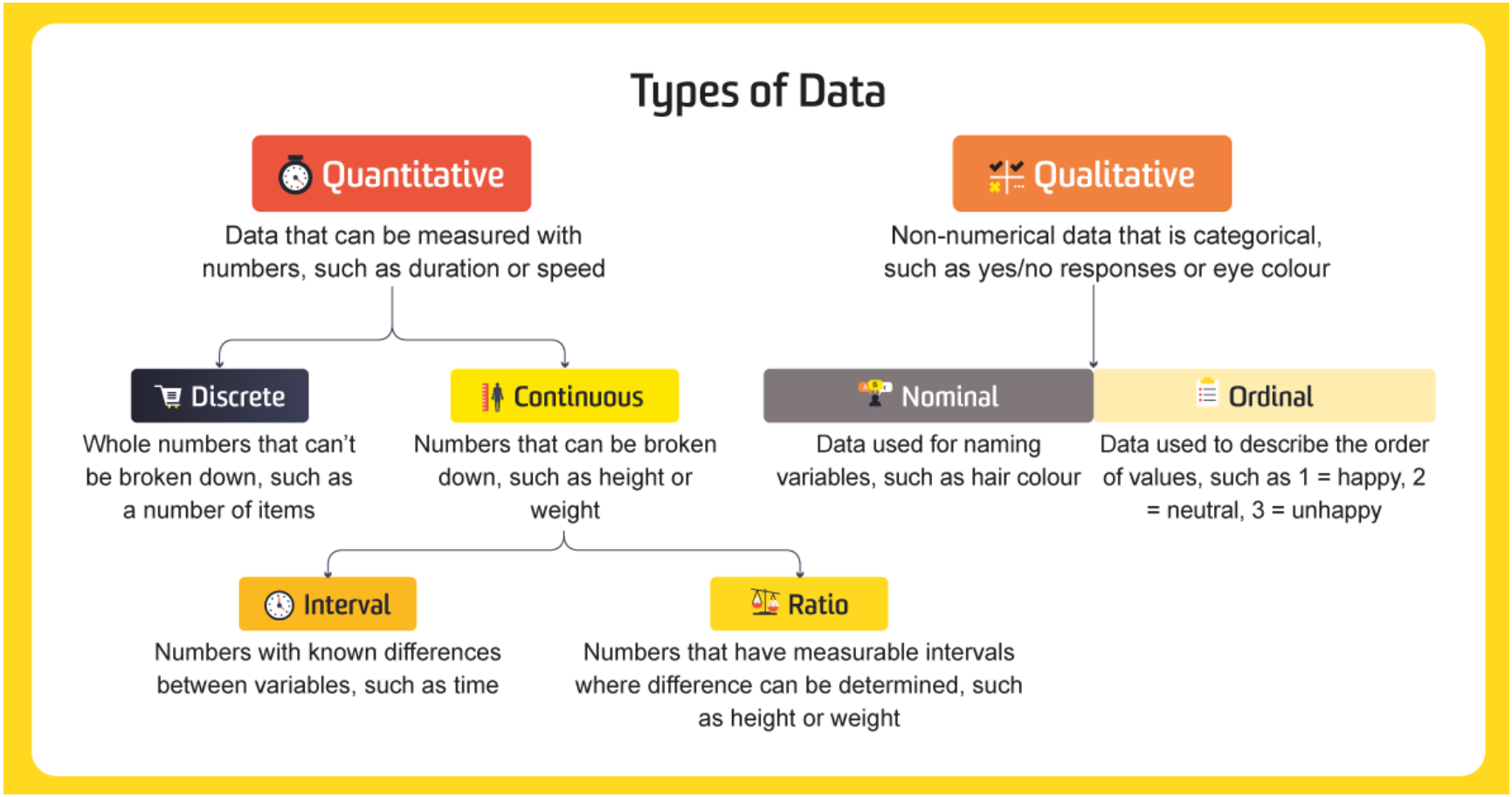
Tipos de variables

- En un EDA, es fundamental identificar los **tipos de variables** a analizar.
- La mayor parte de las variables de un conjunto de datos se divide en dos grupos:
  - **Numéricas**: Tienen un sentido de medida, por ejemplo, edad, altura, peso, tensión arterial, frecuencia cardiaca, temperatura, etc. También se llaman cuantitativas.
  - **Categóricas**: Representan características, por ejemplo, sexo, estado civil, etc. También se llaman cualitativas.

Tipos de variables

- **Numéricas**
  - **Continuas**: Puede tener un número infinito de valores numéricos dentro de un rango específico, por ejemplo, altura, peso, etc.
    - Intervalo
    - Ratio
  - **Discretas**: Sus valores se pueden contar, por ejemplo, número de caras en 100 tiradas de una moneda.
- **Categóricas**
  - Ordinal
  - Nominal

Tipos de variables



Tipos de EDA

- Los **tipos** de EDA se pueden clasificar en:
  - **Univariante no gráfico**: descripción no gráfica de una variable.
  - **Univariante gráfico**: descripción gráfica de una variable.
  - **Multivariante no gráfico**: descripción no gráfica de un conjunto de variables.
  - **Multivariante gráfico**: descripción gráfica de un conjunto de variables.

Tipos de EDA

- **Univariante no gráfico**
  - **Categóricas**
    - Tabla con conteo, proporción y porcentaje de cada categoría para la variable a estudiar.

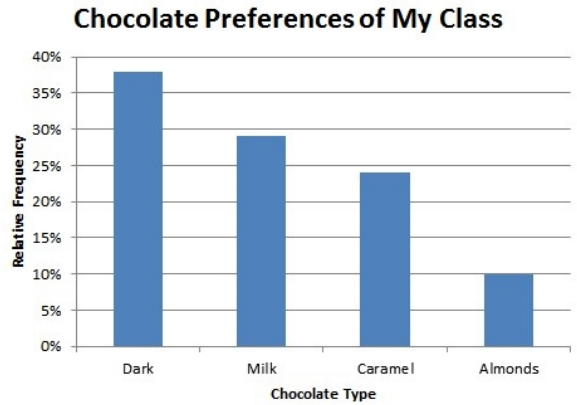
Statistic/College	H&SS	MCS	SCS	other	Total
Count	5	6	4	5	20
Proportion	0.25	0.30	0.20	0.25	1.00
Percent	25%	30%	20%	25%	100%

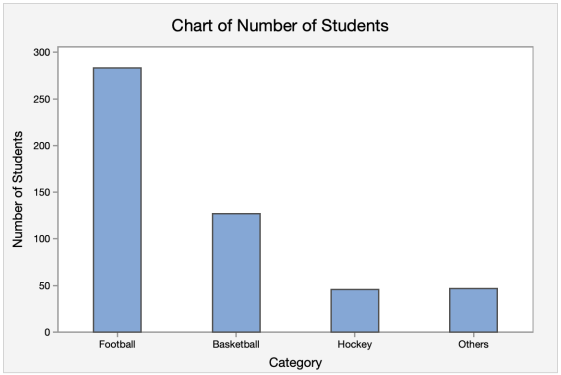
Tipos de EDA

- **Univariante no gráfico**
  - **Numéricas**
    - Tendencia central:
      - Media
      - Mediana
      - Moda
    - Dispersión
      - Varianza
      - Desviación estándar
      - Rango de intercuartil
    - Asimetría y curtosis

Tipos de EDA

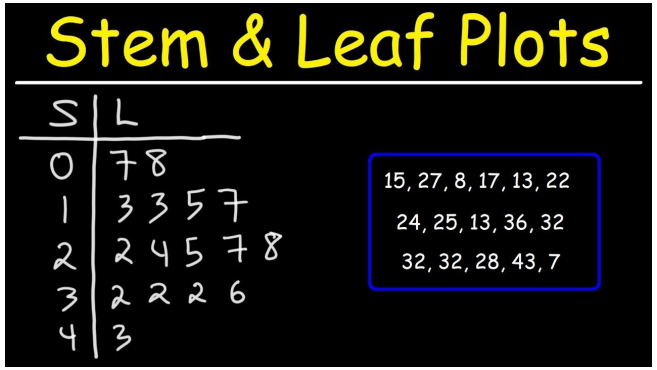
- **Univariante gráfico**
  - **Categóricas**
    - *Bar Chart* (a veces lo llaman histograma también para categóricas)

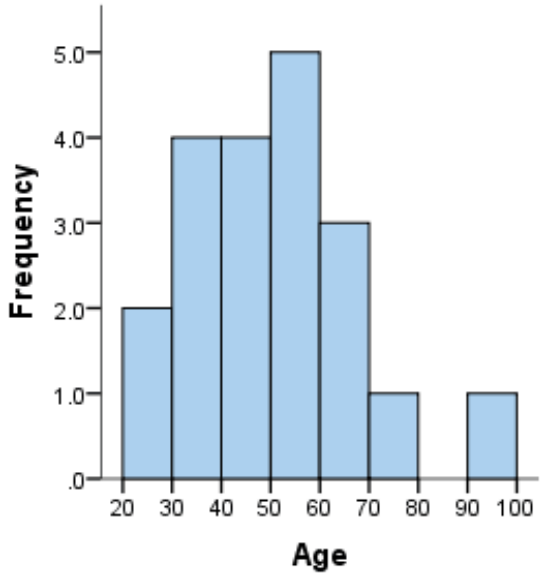


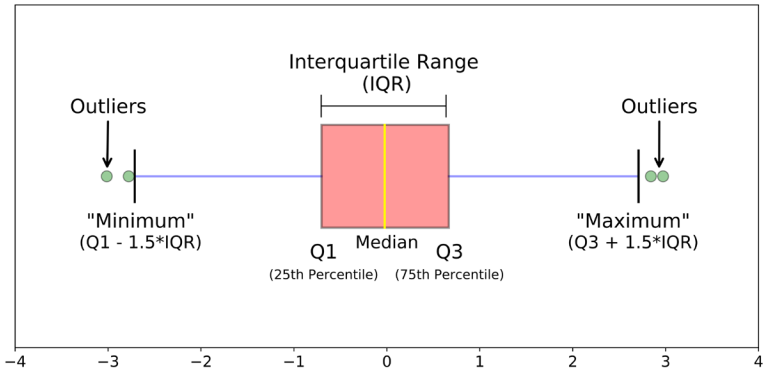


Tipos de EDA

- **Univariante gráfico**
  - **Numéricas**
    - Histograma
    - *Stem-and-leaf* plot
    - *Boxplot*







Tipos de EDA

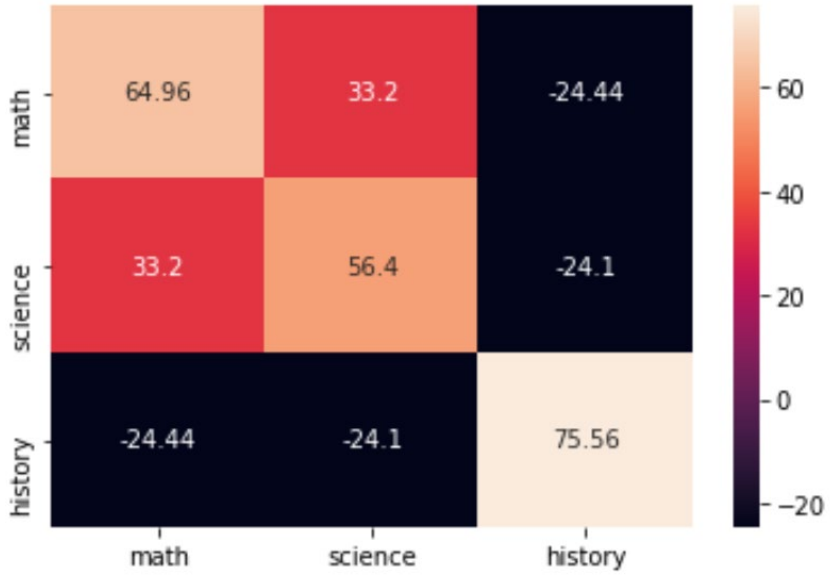
- **Multivariante no gráfico**
  - **Categóricas**
    - Tabla cruzada

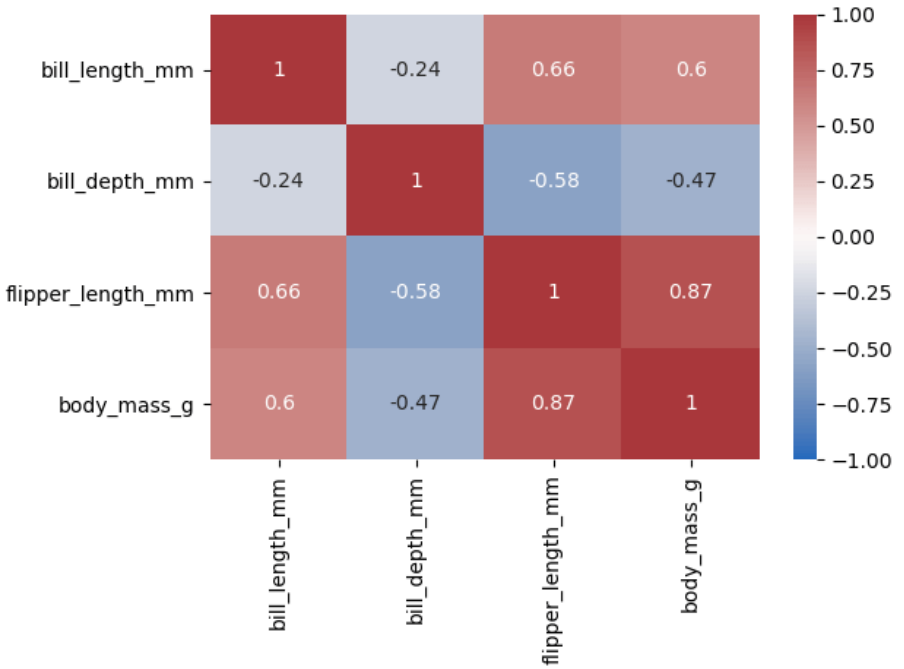
Subject ID	Age Group	Sex
GW	young	F
JA	middle	F
TJ	young	M
JMA	young	M
JMO	middle	F
JQA	old	F
AJ	old	F
MVB	young	M
WHH	old	F
JT	young	F
JKP	middle	M

Age Group / Sex	Female	Male	Total
young	2	3	5
middle	2	1	3
old	3	0	3
Total	7	4	11

Tipos de EDA

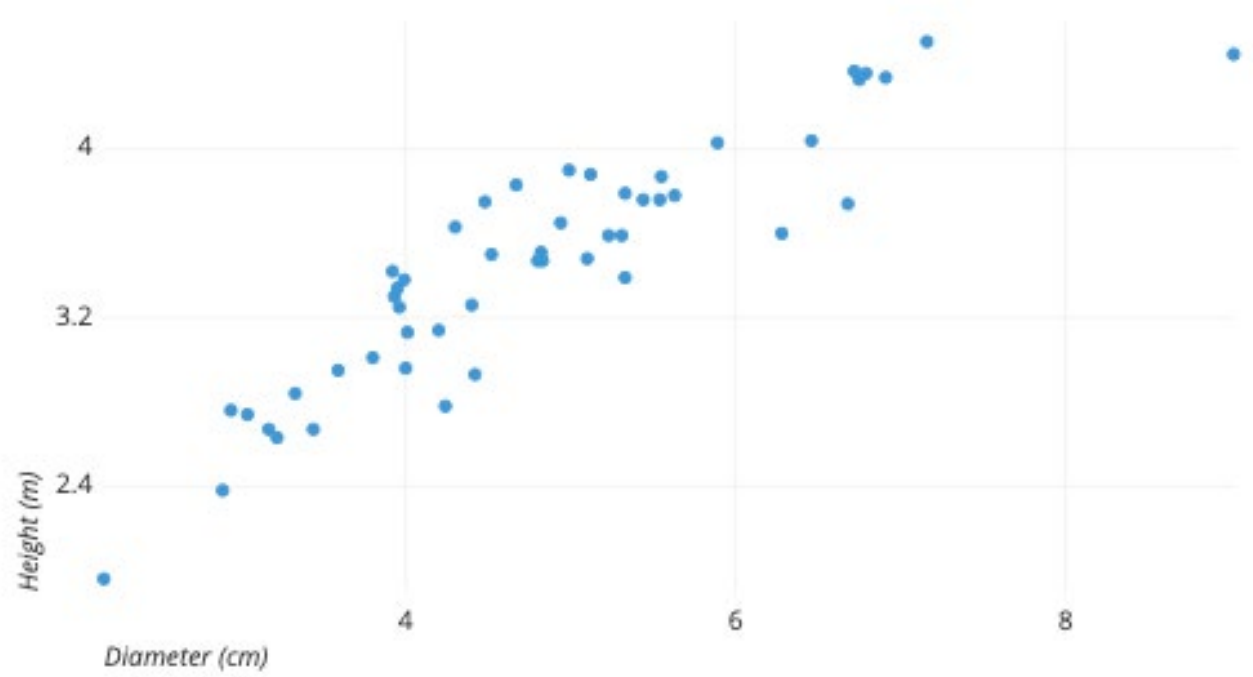
- **Multivariante no gráfico**
  - **Numéricas**
    - Tabla de covarianzas
    - Tabla de correlaciones





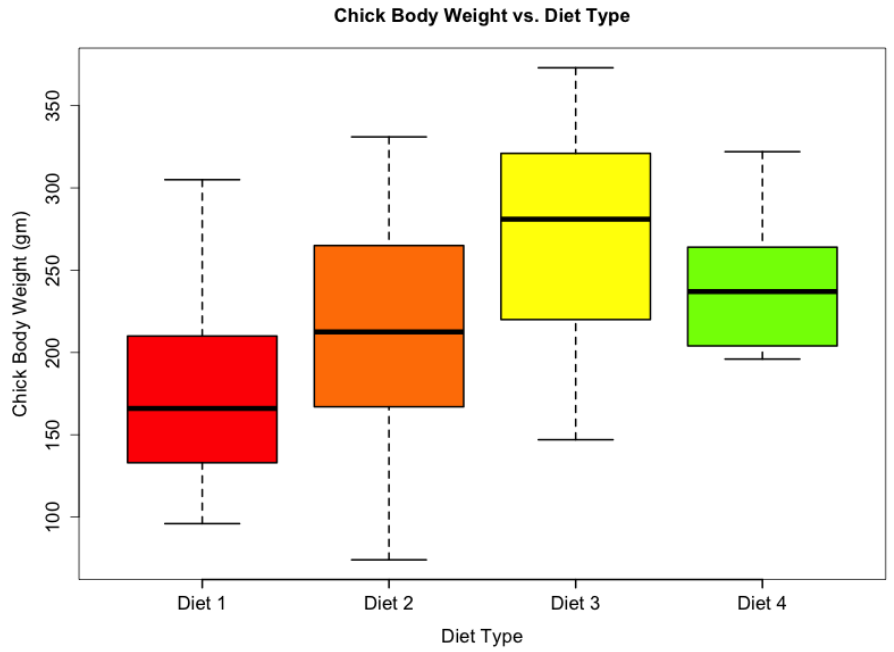
Tipos de EDA

- **Multivariante gráfico**
  - *Scatterplot*: 2 variables numéricas



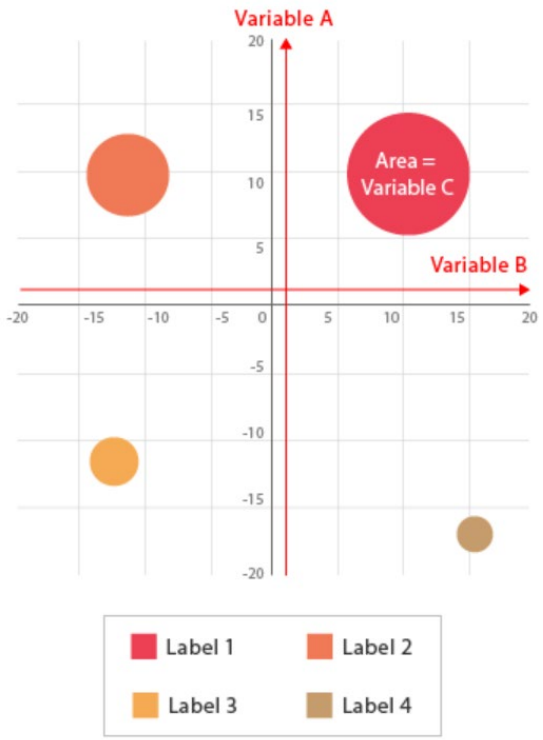
Tipos de EDA

- **Multivariante gráfico**
  - *Side-by-side boxplot*: 1 variable categórica y 1 variable numérica



Tipos de EDA

- **Multivariante gráfico**
  - *Bubble chart*: 1 variable categórica y 2 (o 3) variables numéricas



Conjunto de datos (1)

- *peruvian_blood_pressures.csv*
- Conjunto de datos de personas peruanas que se han trasladado de zonas rurales de gran altitud a zonas urbanas de menor altitud.
- Las variables pueden estar relacionadas con la presión arterial.

	Age	Years	Weight	Height	Chin	Forearm	Calf	Pulse	Systol	Diastol
0	21	1	71.0	1629	8.0	7.0	12.7	88	170	76
1	22	6	56.5	1569	3.3	5.0	8.0	64	120	60
2	24	5	56.0	1561	3.3	1.3	4.3	68	125	75
3	24	1	61.0	1619	3.7	3.0	4.3	52	148	120
4	25	1	65.0	1566	9.0	12.7	20.7	72	140	78

Conjunto de datos (1)

- Las variables de este conjunto de datos son:
  - *Age*: edad (años)
  - *Years*: años en zona urbana
  - *Weight*: peso (kg)
  - *Height*: altura (mm)
  - *Chin:* pliegue de la barbilla (mm)
  - *Forearm:* pliegue cutáneo del antebrazo (mm)
  - *Calf*: pliegue cutáneo de la pantorrilla (mm)
  - *Pulse*: frecuencia del pulso en reposo (bpm)
  - *Systol*: presión arterial sistólica (mmHg)
  - *Diastol*: presión arterial diastólica (mmHg)

Conjunto de datos (2)

- *sourth_africa_chd.csv*
- Conjunto de datos de varones de una región de Sudáfrica con alto riesgo de cardiopatías.
- Las variables pueden estar relacionadas con la presencia de cardiopatías.

	sbp	tobacco	ldl	adiposity	famhist	typea	obesity	alcohol	age	chd
0	160	12.00	5.73	23.11	Present	49	25.30	97.20	52	Si
1	144	0.01	4.41	28.61	Absent	55	28.87	2.06	63	Si
2	118	0.08	3.48	32.28	Present	52	29.14	3.81	46	No
3	170	7.50	6.41	38.03	Present	51	31.99	24.26	58	Si
4	134	13.60	3.50	27.78	Present	60	25.99	57.34	49	Si

Conjunto de datos (2)

- Las variables de este conjunto de datos son:
  - *sbp*: presión arterial sistólica (mmHg)
  - *tobacco*: tabaco acumulado (kg)
  - *ldl*: colesterol de lipoproteínas de baja densidad
  - *adiposity*: Medida de adiposidad
  - *famhist*: antecedentes familiares de enfermedades cardíacas
  - *typeA*: Comportamiento de personalidad propenso a coronarias de tipo A
  - *obesity*: Medida de obesidad
  - *alcohol*: consumo actual de alcohol
  - *age*: edad
  - *chd*: cardiopatía coronaria

Bibliografía

- Mukhiya & Ahmed. *Hands-On Exploratory Data Analysis with Python* (2020), Packt Publishing.
- Seltman. *Experimental Design and Analysis* (2018) https://www.stat.cmu.edu/~hseltman/309/Book/Book.pdf