DEPARTAMENTO
DE COMPUTACION

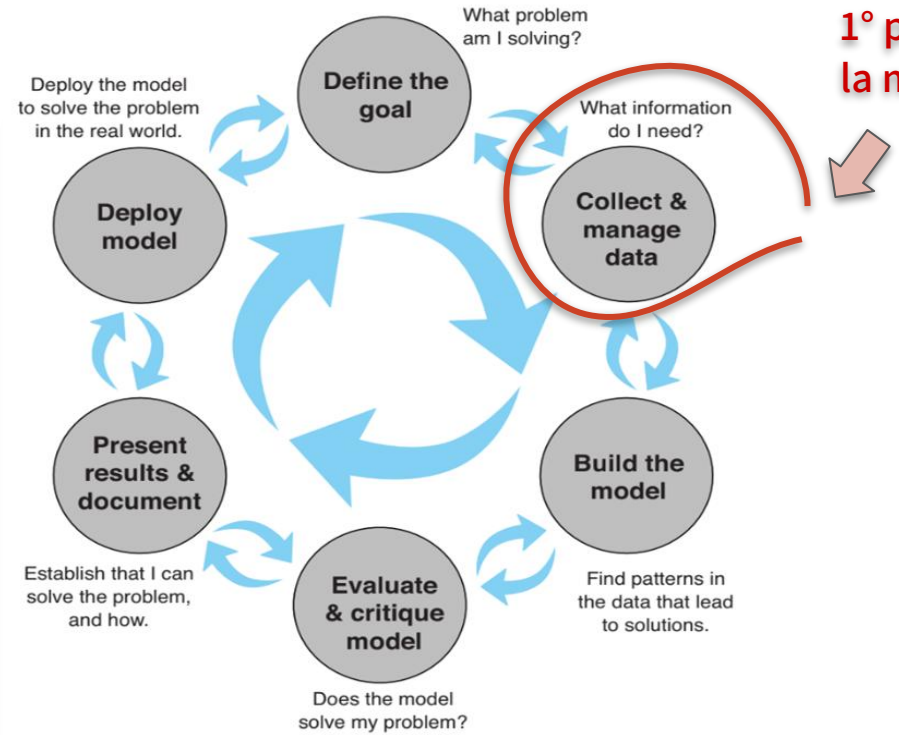
Facultad de Ciencias Exactas y Naturales - UBA

Laboratorio de datos

Visualización y Análisis Exploratorio de Datos

Verano 2026

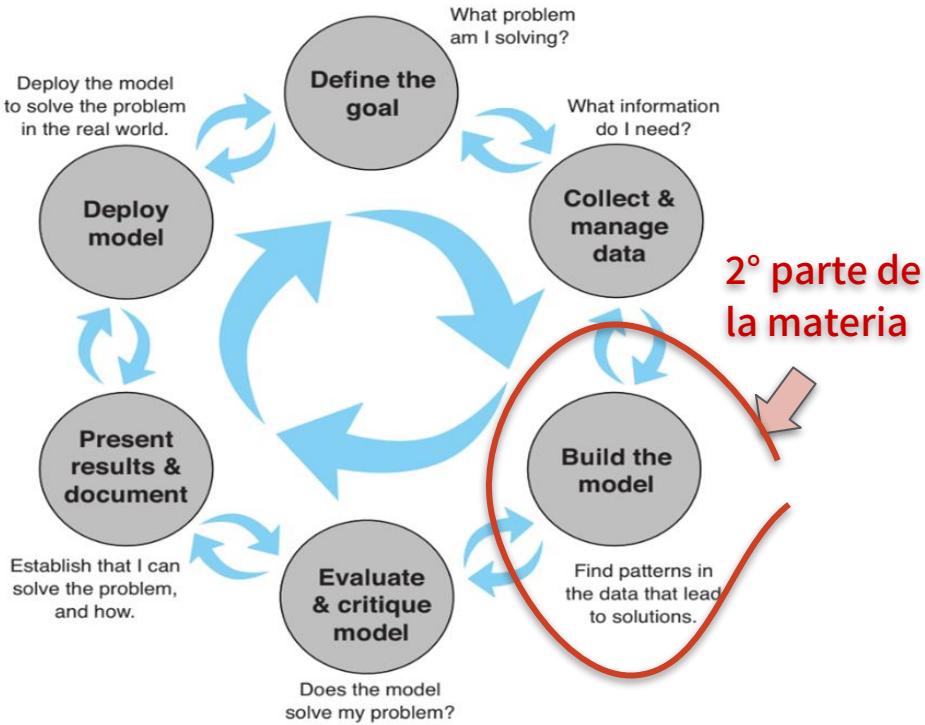
Hasta ahora vimos...



1° parte de la materia

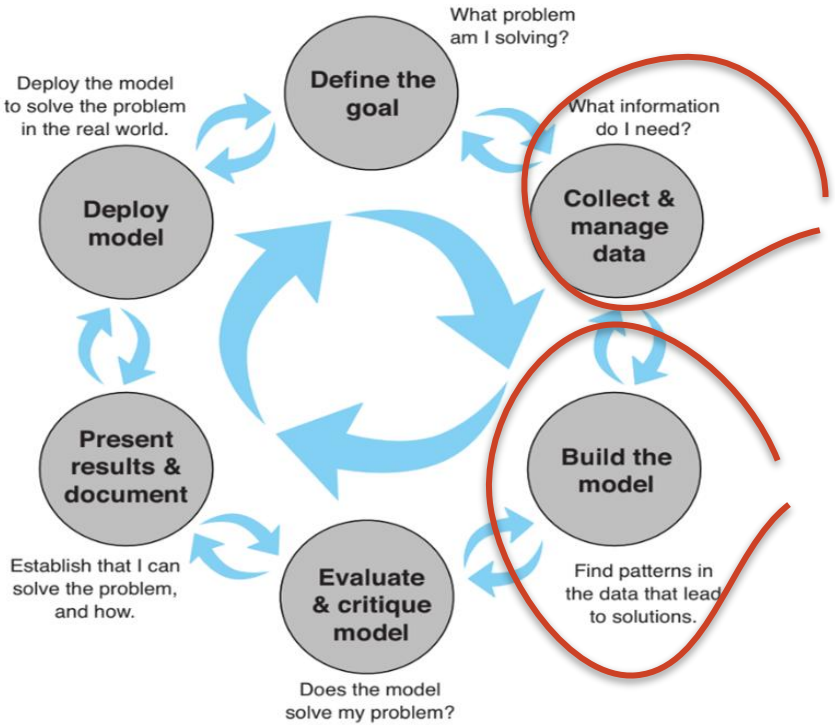
- Lenguaje de programación (Python)
- Modelado conceptual de los datos (DER)
- Representación de los datos (modelo relacional)
- Formas de consultar los datos (AR/SQL)
- Recomendaciones para el diseño (Normalización)
- Calidad de datos

Clase de hoy



- **Visualización y Exploración de los datos**

Clase de hoy



Administrar el **almacenamiento** de los datos:

- Qué datos me conviene guardar?
- Cuáles atributos definen a una entidad?
- Minimizar redundancias y anomalías
- Restricciones de integridad

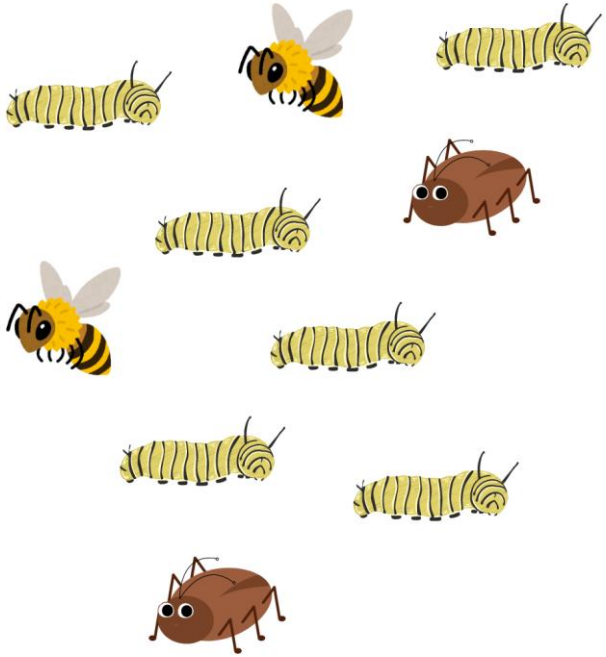
Encontrar **relaciones** entre los datos:

- Edad y altura crecen del mismo modo?
- Las causas de defunción tienen la misma frecuencia en todo el país?
- La frecuencia de las causas depende del rango etario?

¿Qué es el proceso de análisis de datos?

Proceso científico de **transformar** datos en información para tomar mejores decisiones

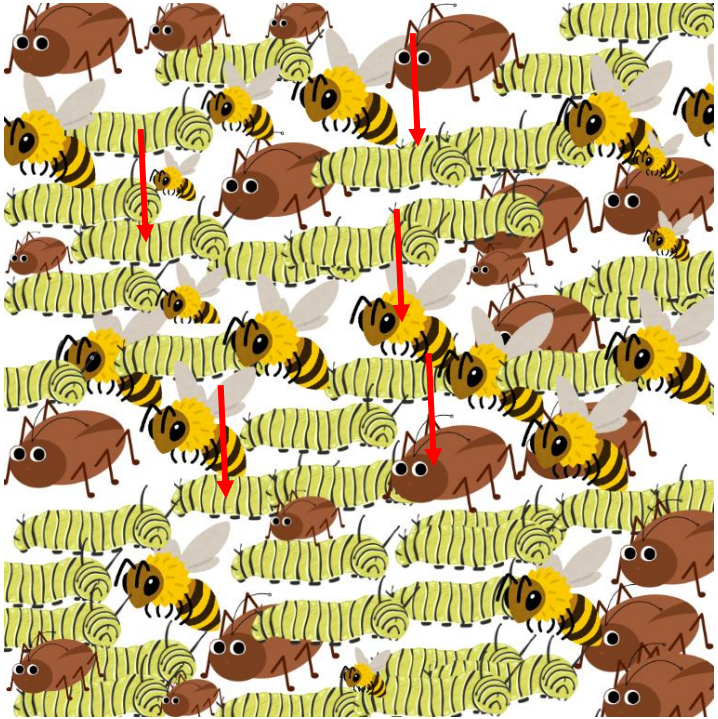
Alcance del análisis de datos



Queremos caracterizar la composición de esta **población** de insectos:

- 2 abejas
- 2 cucarachas
- 6 orugas

Alcance del análisis de datos



Recopilar datos de toda la población sería **costoso**

Vamos a recopilar datos de un **subconjunto** de la población → **Muestra (*sample*)**

Si asumimos que una **muestra** de datos es **representativa de la población**, podemos hacer **generalizaciones** sobre toda la población

Alcance del análisis de datos

Análisis descriptivo. Herramientas que describen lo que ha sucedido.
Ej. Queries, reportes, estadística descriptiva, visualización de datos. En general, resumen los datos existentes o los resultados de análisis predictivos o prescriptivos

Análisis predictivo. Técnicas que utilizan modelos matemáticos construidos a partir de datos pasados para predecir eventos futuros o comprender mejor las relaciones entre variables. Ej. Análisis de regresión, simulaciones computacionales

Análisis prescriptivo. Son modelos matemáticos o lógicos que sugieren una decisión o un curso de acción. Ej. modelos de optimización matemática, evaluación de escenarios, análisis de decisiones

Alcance del análisis de datos

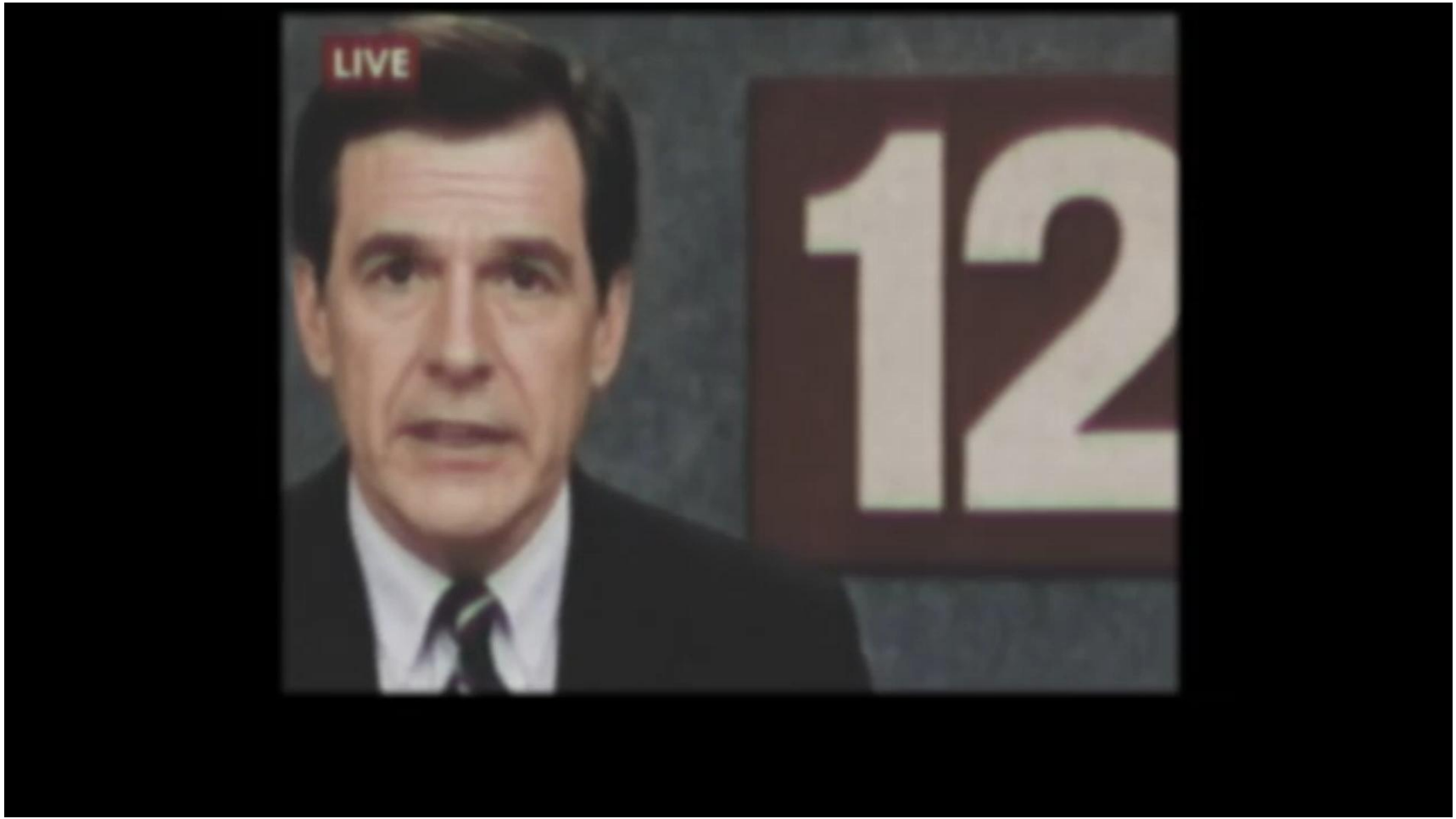
- Análisis descriptivo
- Análisis predictivo
- Análisis prescriptivo

La **visualización** de datos es fundamental para el éxito de los tres tipos de análisis

Para qué visualizar

Al transformar los números y otras piezas de información en gráficos, el contenido se hace más fácil de **entender y usar.**

- resumir información
- encontrar tendencias o patrones
- detectar valores anómalos
- encontrar relaciones entre variables



This is news: https://www.youtube.com/watch?v=SHb-3oIAFTs

Para qué visualizar

Al transformar los números y otras piezas de información en gráficos, el contenido se hace más fácil de **entender y usar.**

- resumir información
- encontrar tendencias o patrones
- detectar valores anómalos
- encontrar relaciones entre variables

Para qué visualizar

Al transformar los números y otras piezas de información en gráficos, el contenido se hace más fácil de **entender y usar.**

- resumir información
- encontrar tendencias o patrones
- detectar valores anómalos
- encontrar relaciones entre variables
- hacerse preguntas, y en consecuencia elaborar hipótesis
- mostrar o reforzar una hipótesis
- mostrar resultados

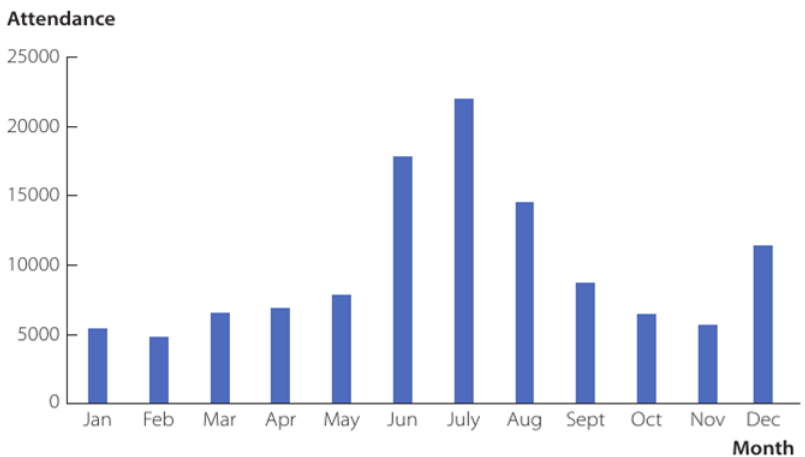
Ejemplo: encontrar tendencias o patrones

TABLE 1.1 Zoo Attendance Data

Month	Jan	Feb	Mar	Apr	May	Jun
Attendance	5422	4878	6586	6943	7876	17843
Month	July	Aug	Sept	Oct	Nov	Dec
Attendance	21967	14542	8751	6454	5677	11422

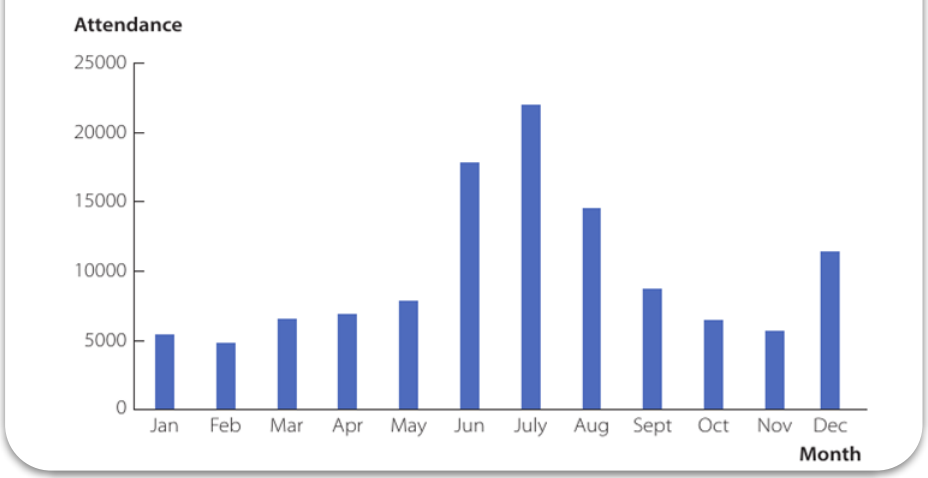
¿Cómo varía la asistencia al zoológico a lo largo del tiempo?

FIGURE 1.1 A Column Chart of Zoo Attendance by Month



- La asistencia aumenta en junio y julio, cuando los niños en edad escolar no asisten a la escuela (vacaciones de verano - hemisferio norte).
- Aumento gradual de la asistencia con la temperatura (desde febrero hasta mayo).

FIGURE 1.1 A Column Chart of Zoo Attendance by Month



- La asistencia en diciembre no sigue estos patrones → El zoo cuenta con el “Festival de las Luces” que se extiende desde finales de noviembre hasta principios de enero. También corresponde al periodo de vacaciones de invierno de los niños en edad escolar

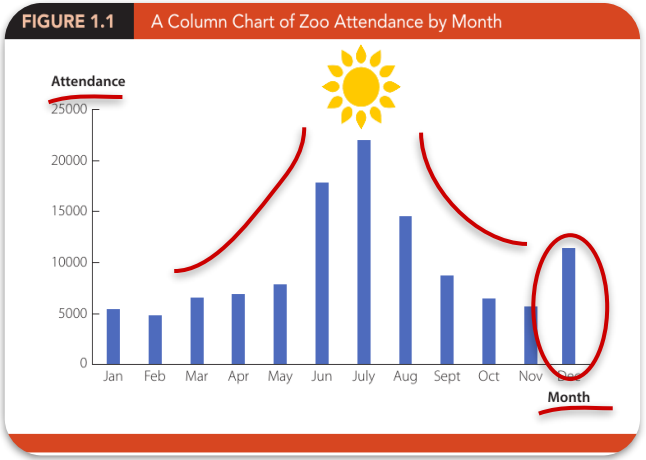
Visualización de datos según la finalidad

La **exploración** visual de datos es una parte crucial del análisis descriptivo.

La exploración de datos permite:

- Identificar patrones
- Reconocer anomalias e irregularidades
- Caracterizar la relación entre variables

La **visualización** nos da mayor capacidad para **detectar patrones, anomalías y relaciones** entre variables que al hacerlo mirando simplemente los datos crudos



Mostrar o reforzar una hipótesis

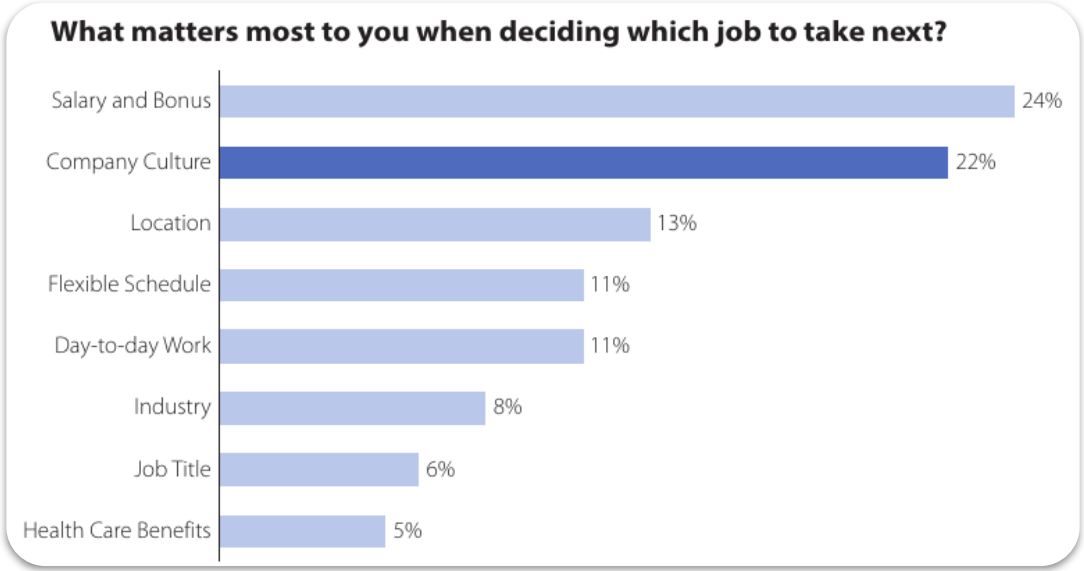
Queremos mostrar que la “Cultura de la compañía” es uno de los factores más importantes a la hora de buscar trabajo

¿Resulta obvio?

Factor	Porcentaje
Flexible Schedule	11,00%
Location	13,00%
Salary and Bonus	24,00%
Job Title	6,00%
Health Benefit Benefits	5,00%
Industry	8,00%
Company Culture	22,00%
Day-to-day Work	11,00%

Queremos mostrar que la “Cultura de la compañía” es uno de los factores más importantes a la hora de buscar trabajo

¿Resulta obvio?

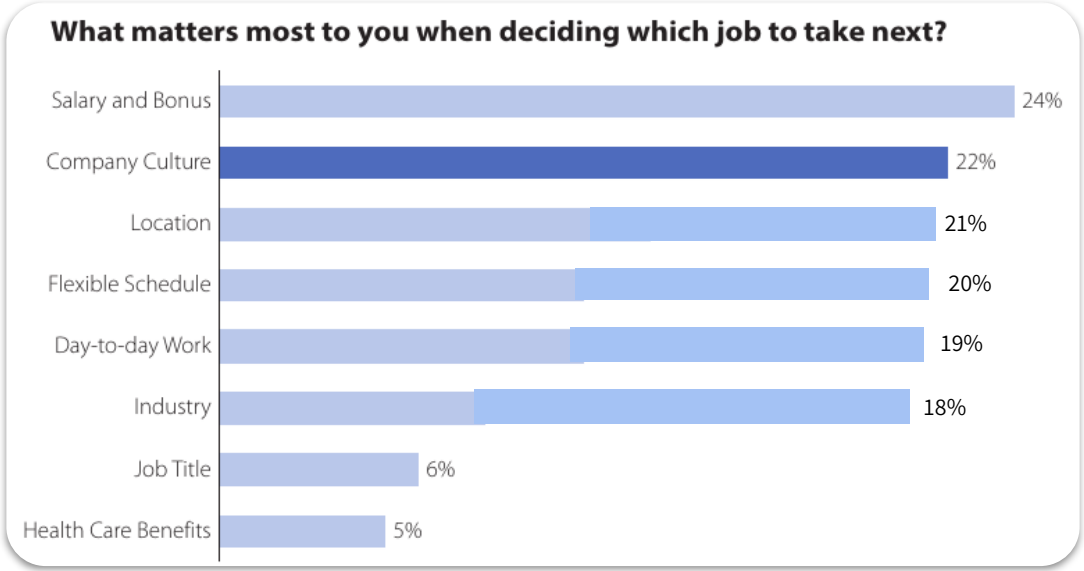


La **visualización** de datos es útil para **comunicar** a la audiencia y garantizar que **comprenda** y se concentre en el mensaje deseado.

Visualización de datos según la finalidad: explicación

Queremos mostrar que la “Cultura de la compañía” es uno de los factores más importantes a la hora de buscar trabajo

¿Resulta obvio?



En este caso, a pesar de que el ranking es el mismo y la conclusión sigue siendo estrictamente válida, estaríamos introduciendo un sesgo cuando en realidad tiene casi el mismo peso que otros 4 factores!

Tipos de datos

El tipo de gráfico a realizar depende de las características de los datos con los que se cuenta:

- Nominales - categorías
- Ordinales - escala discretas, rankings
- Cuantitativos - magnitudes
  - Proporciones - comparaciones de magnitudes
- Transversales
- Series Temporales

Tipos de datos: cuantitativos vs categóricos

¿Qué tipo de variable es *Length?*

Variable cuantitativa

Sex	Length	Diameter	Height	Weight	Shucked Weight	Viscera Weight	Shell Weight	Age
F	16875	1275	375	3392017675	134660125	8618248	10942907	11
M	11	8875	3125	1353688625	3742134	231048425	5386405	9
F	15	1175	475	3216250775	13947954	735669525	956795625	10
M	13125	9625	25	1450076925	6973977	284912475	412485225	8
I	9875	675	25	846232575	409650275	17293195	2324659	5
I	13125	1	35	1971707725	681805475	453592	71724235	10
F	1325	1075	375	210069795	92135875	525883225	5556502	9
I	11375	9375	3125	151102835	66054335	3005047	52446575	8
I	1075	85	275	1033339275	45075705	242388225	29766975	7
M	1325	11	5125	236718325	907184	616601625	69456275	14
M	15375	13125	3875	3224755625	104042665	6690482	10489315	20
F	13125	10125	4	18653971	752679225	318931875	63786375	12

Variables cuantitativas:

- Permiten indicar una magnitud.
- Se les puede aplicar operaciones aritméticas (+, -, *, %, etc.).

Tipos de datos: cuantitativos vs categóricos

¿Qué tipo de variable es *Sex?*

Variable categórica

Sex	Length	Diameter	Height	Weight	Shucked Weight	Viscera Weight	Shell Weight	Age
F	16875	1275	375	3392017675	134660125	8618248	10942907	11
M	11	8875	3125	1353688625	3742134	231048425	5386405	9
F	15	1175	475	3216250775	13947954	735669525	956795625	10
M	13125	9625	25	1450076925	6973977	284912475	412485225	8
I	9875	675	25	846232575	409650275	17293195	2324659	5
I	13125	1	35	1971707725	681805475	453592	71724235	10
F	1325	1075	375	210069795	92135875	525883225	5556502	9
I	11375	9375	3125	151102835	66054335	3005047	52446575	8
I	1075	85	275	1033339275	45075705	242388225	29766975	7
M	1325	11	5125	236718325	907184	616601625	69456275	14
M	15375	13125	3875	3224755625	104042665	6690482	10489315	20
F	13125	10125	4	18653971	752679225	318931875	63786375	12

Variables categóricas:

- Permiten identificar ítems similares mediante **etiquetas** o nombres.
- No se pueden realizar operaciones aritméticas con datos categóricos. Sin embargo, se pueden **sintetizar** los datos categóricos contando el número de observaciones o calculando las **proporciones** de las observaciones de cada categoría.

Tipos de datos: transversales vs series temporales

Los datos de esta tabla ¿Son transversales o corresponden a una serie temporal?

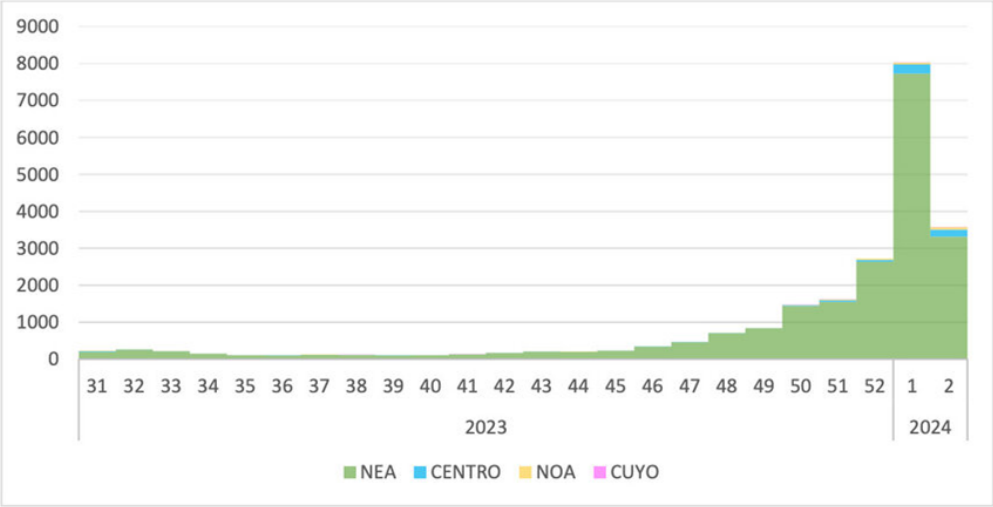
Sex	Length	Diameter	Height	Weight	Shucked Weight	Viscera Weight	Shell Weight	Age
F	16875	1275	375	3392017675	134660125	8618248	10942907	11
M	11	8875	3125	1353688625	3742134	231048425	5386405	9
F	15	1175	475	3216250775	13947954	735669525	956795625	10
M	13125	9625	25	1450076925	6973977	284912475	412485225	8
I	9875	675	25	846232575	409650275	17293195	2324659	5
I	13125	1	35	1971707725	681805475	453592	71724235	10
F	1325	1075	375	210069795	92135875	525883225	5556502	9
I	11375	9375	3125	151102835	66054335	3005047	52446575	8
I	1075	85	275	1033339275	45075705	242388225	29766975	7
M	1325	11	5125	236718325	907184	616601625	69456275	14
M	15375	13125	3875	3224755625	104042665	6690482	10489315	20
F	13125	10125	4	18653971	752679225	318931875	63786375	12

Datos transversales (Cross-Sectional Data)
Son datos que corresponden al mismo momento o aprox. del mismo tiempo.

Tipos de datos: transversales vs series temporales

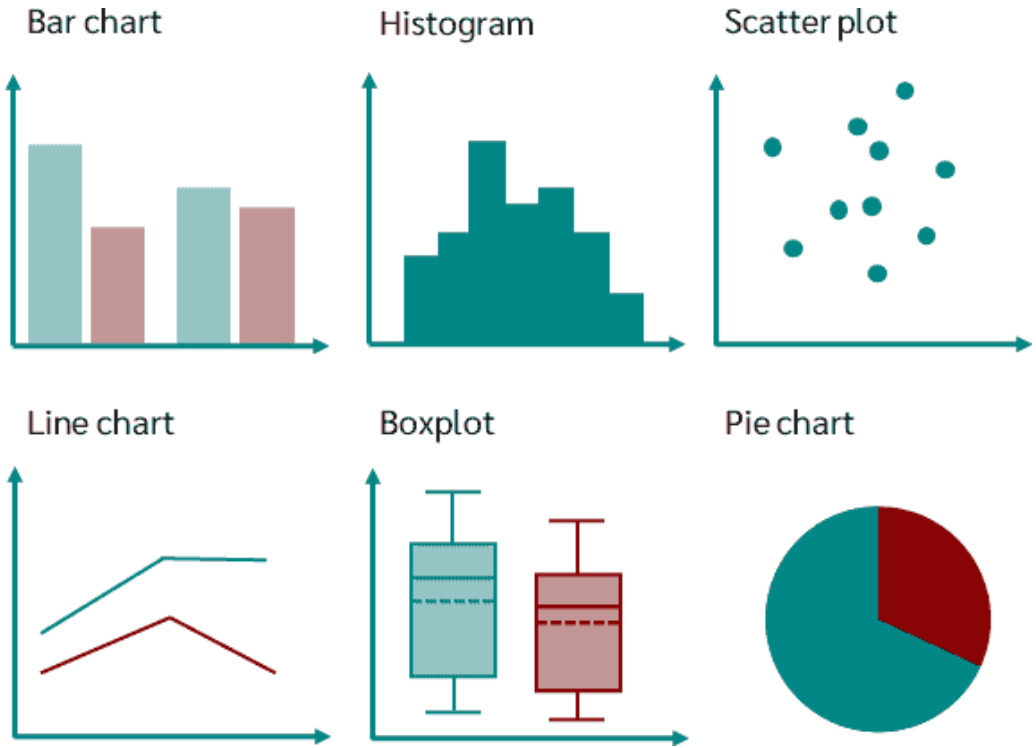
Los datos de esta figura ¿Son transversales o corresponden a una serie temporal?

Gráfico 1. Casos de Dengue sin antecedentes de viaje por semana epidemiológica según región. SE 31/2023 a SE 2/2024 (n=22.466). Argentina.

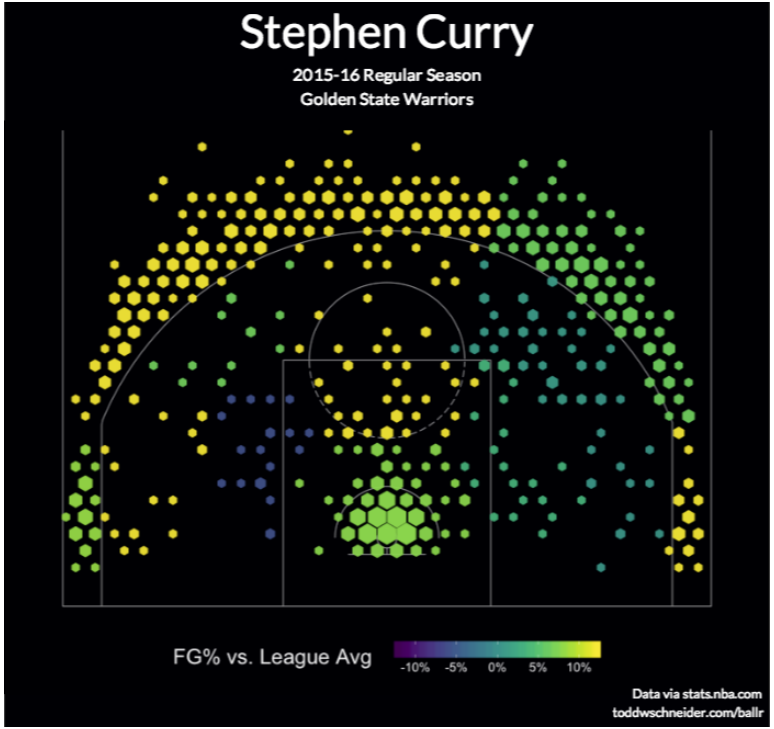


Series de tiempo (Time Series Data)
Son datos recopilados en varios puntos de tiempo (minutos, horas, días, meses, años, etc.).
Estos gráficos ayudan a los analistas a **comprender** lo que sucedió en **el pasado**, identificar **tendencias** a lo largo del tiempo y **proyectar** niveles **futuros** para la serie temporal.
Ej. Los gráficos de datos de series de tiempo se encuentran con frecuencia en publicaciones comerciales, económicas y científicas.

Visualización de datos: ejemplos comunes



Visualización de datos: ejemplos menos comunes



Visualización de datos: ejemplos menos comunes

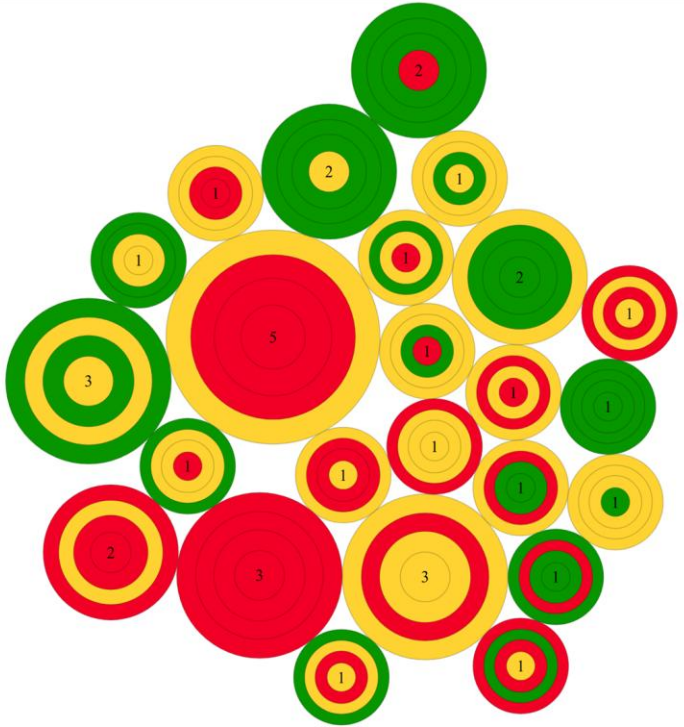
Medición de escuelas primarias

Desde adentro hacia afuera:

- Cognición y lenguaje
- Habilidades cotidianas
- Adaptación comportamental
- Soporte familiar

Cada diferente patrón de aros de colores es una combinación específica de cuatro valores

(López y Rosenfeld)



Selección del tipo de gráfico

1. Objetivo:
   a. Explorar. Va a depender de la pregunta a responder y qué se espera de los datos
   b. Explicar. Va a depender del mensaje que se quiere dar

2. Tipos de datos a graficar:
   a. Variables cuantitativas/cualitativas
   b. Datos Transversales vs. Series Temporales

3. Otros objetivos:
   a. Ranking. Conocer el orden relativo de los elementos.
   b. Correlación/Relación. Comprender cómo dos variables se relacionan entre sí. Ej. relación entre la temperatura mínima promedio y las nevadas anuales promedio en varias ciudades de Argentina.
   c. Distribución. Saber cómo se dispersan los ítems. Ej. Cantidad de llamadas que recibe un call center a lo largo de un día.
   d. Composición. Entender cómo se constituye una cierta entidad. Ej. Voto de las últimas elecciones.

Selección del tipo y formato de gráfico

Objetivo

- Qué preguntas queremos responder
- Qué se espera de los datos
- Qué queremos mostrar
- Qué queremos enfatizar o resaltar
- Quién va a ver el gráfico
- Para qué se va a usar el gráfico
- Estilo que queremos utilizar

Tipos de datos a graficar

- Cuántas variables tenemos que representar
- Qué tipos variables tenemos que representar
- Tipos de interacción entre las variables
- Si los datos son transversales, temporales, u otros (espaciales…)

Nuestros primeros gráficos

Dataset de vinos

type	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	pH	sulphates	alcohol	quality
white	8.4	0.17	0.31	6.7	0.038	29	132	3.1	0.32	10.6	7
white	6	0.18	0.31	1.4	0.036	14	75	3.34	0.58	11.1	8
white	8.6	0.36	0.26	11.1	0.03	43.5	171	3.03	0.49	12	5
white	6.9	0.4	0.17	12.9	0.033	59	186	3.08	0.49	9.4	5
red	6.8	0.785	0	2.4	0.104	14	30	3.52	0.55	10.7	6
red	10.8	0.29	0.42	1.6	0.084	19	27	3.28	0.73	11.9	6
white	7.1	0.21	0.32	2.2	0.037	28	141	3.2	0.57	10	7
white	6.1	0.17	0.21	1.9	0.09	44	130	3.07	0.41	9.7	5
white	9.2	0.28	0.46	3.2	0.058	39	133	3.14	0.58	9.5	5
red	11.5	0.59	0.59	2.6	0.087	13	49	3.18	0.65	11	6

Nuestros primeros gráficos

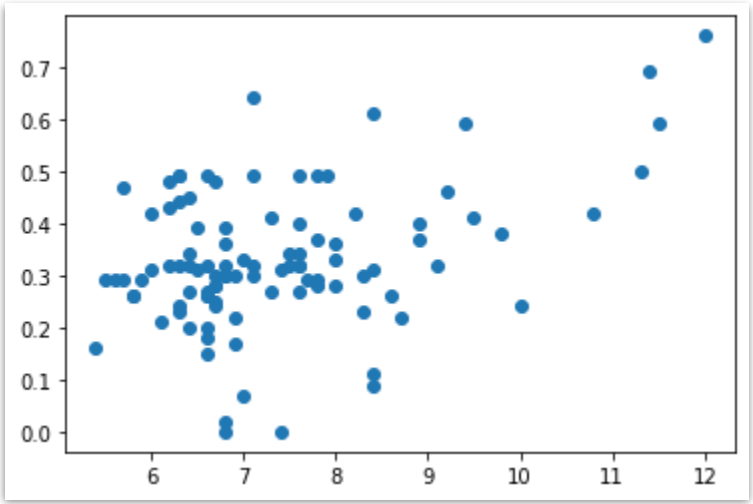
```
# Importar Bibliotecas
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt # Para graficar series multiples

# Carpeta donde se encuentran los archivos a utilizar
carpeta = 'data/'
```

Leemos el *dataset*

```
wine = pd.read_csv(carpeta+"wine.csv", sep = ";")
# con sep indicamos que el separador es ;
```

Gráfico de dispersión/gráfico de puntos/*scatter plot*

```
# Genera el grafico que relaciona la acidez (no volatil) y el contenido de
# acido citrico de cada vino
plt.scatter(data = wine, x='fixed acidity', y='citric acid')
# plt.scatter(wine['fixed acidity'], wine['citric acid’]) #otra manera
```



Figuras de Matplotlib

`fig, ax = plt.subplots()` → devuelve una tupla!

```
In[7]:   plt.subplots()
Out[7]:  (<Figure size 864x576 with 1 Axes>, <Axes: >)
```

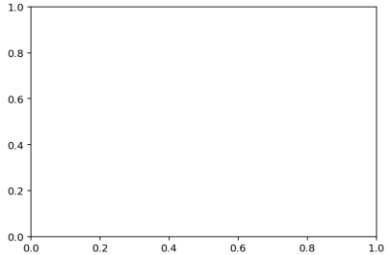


Fig: contenedor principal de todo el gráfico. Puede contener uno o varios ejes (en este caso contiene uno)

- Vamos a usar fig para hacer ajustes globales sobre toda la figura (tamaño, guardado de la imágen, etc)

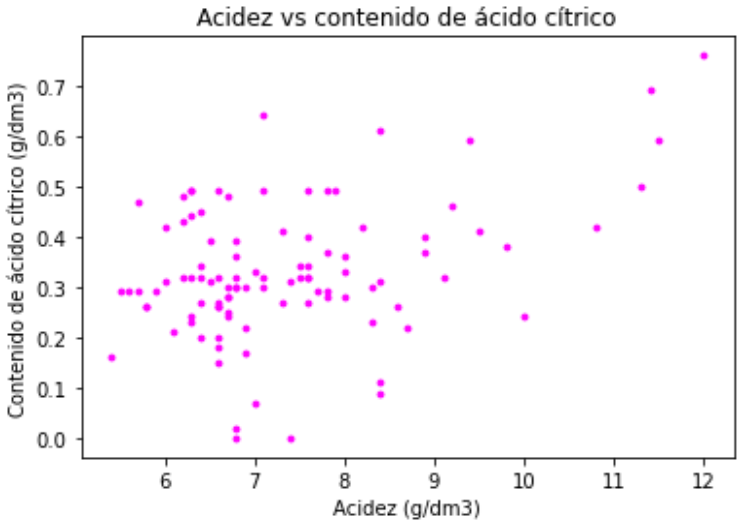
Ax: Ejes, gráficos dentro de cada figura.

- Vamos a usar los atributos de ax para ajustar los elementos de cada gráfico, como etiquetas, lineas, colores, etc.

Gráfico de dispersión/gráfico de puntos/*scatter plot*

`fig, ax = plt.subplots()` → devuelve una tupla!

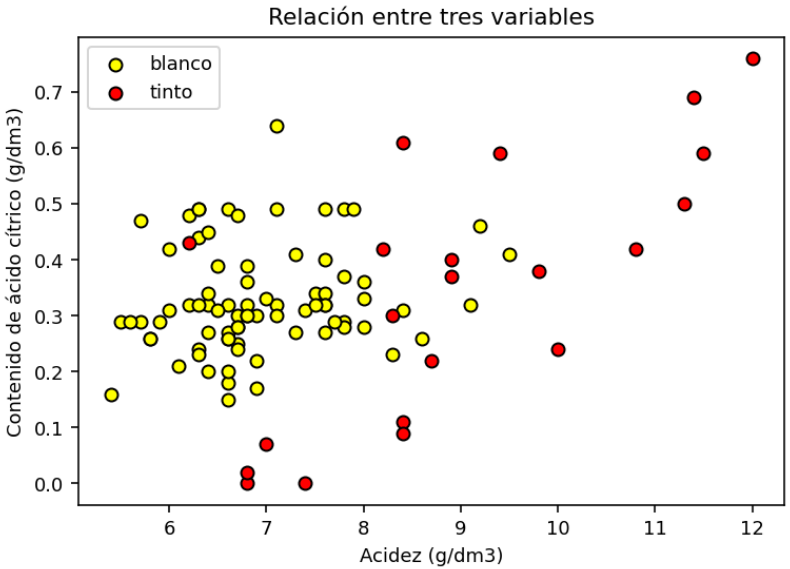
```
ax.scatter(data = wine,
           x='fixed acidity',
           y='citric acid',
           s=8,                        # Tamano de los puntos
           color='magenta')              # Color de los puntos

ax.set_title('Acidez vs contenido de ácido cítrico’)  # Titulo del gráfico
ax.set_xlabel('Acidez (g/dm3)', fontsize='medium’)    # Nombre eje X
ax.set_ylabel('Contenido de ácido cítrico (g/dm3)',
              fontsize='medium')                      # Nombre eje Y
```



¿De qué tipo son las variables graficadas?
¿Cualitativas o cuantitativas?

Scatter Plot con color

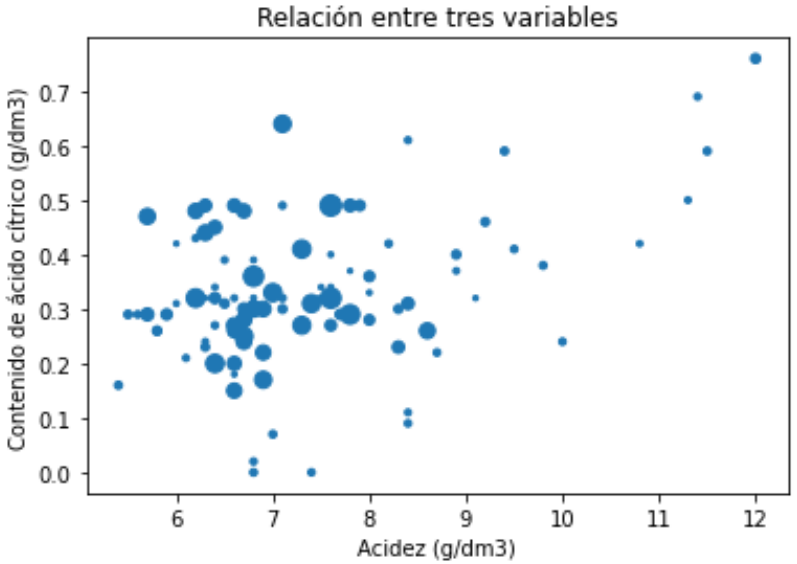
```
# Genera el grafico que relaciona tres variables en simultaneo
fig, ax = plt.subplots()

wine_blanco = wine[wine['type'] == 'white']
wine_tinto  = wine[wine['type'] == 'red']

ax.scatter(data=wine_blanco, x='fixed acidity',
           y='citric acid', c='yellow', edgecolor='k', label='blanco')

ax.scatter(data=wine_tinto, x='fixed acidity',
           y='citric acid', c='red', edgecolor='k', label='tinto')


ax.set_title('Relación entre tres variables')
ax.set_xlabel('Acidez (g/dm3)', fontsize='medium')
ax.set_ylabel('Contenido de ácido cítrico (g/dm3)',
              fontsize='medium')
ax.legend()

del wine_blanco, wine_tinto
```



¿De qué tipo son las variables graficada
¿Cualitativas o cuantitativas?

Gráfico de globos/burbujas (*Bubble chart*)

```
# Genera el grafico que relaciona tres variables en simultaneo
# (mejorando la informacion mostrada)
fig, ax = plt.subplots()

tamanoBurbuja = 5
# Cuanto queremos modificar el tamaño de cada burbuja

ax.scatter(data=wine, x='fixed acidity',
           y='citric acid', s=wine['residual sugar']*tamanoBurbuja)

ax.set_title('Relación entre tres variables')
ax.set_xlabel('Acidez (g/dm3)', fontsize='medium')
ax.set_ylabel('Contenido de ácido cítrico (g/dm3)',
              fontsize='medium')


# remueve la variable temporal tamanoBurbuja que ya no utilizaremos
del(tamanoBurbuja)
```



¿De qué tipo son las variables graficadas?
¿Cualitativas o cuantitativas?

Scatterplot - características

- Representa -al menos- dos variables, en los dos ejes.
  - Estas dos variables son de tipo numérico.
- Pueden sumarse otras variables mediante el uso del color, tipo o tamaño de los marcadores.
  - Estas otras variables pueden ser numéricas, categóricas, ordinales, etc.
- Puede ser útil para entender correlación entre variables.
- Puede ser útil para entender la distribución de valores para cada una de las variables.

Ejercicio

Con el dataset de vinos

¿Existe o no alguna **relación** entre el pH de los vinos (*pH*) y alguna de las otras variables? Mostrarlo gráficamente

Discutir con el resto de la clase:

- ¿Cuál fue su objetivo: Explorar, Explicar, Otro?
- ¿Qué tipos de variables estaban en juego?
- ¿Mejoró alguna característica del gráfico para cumplir con el objetivo?

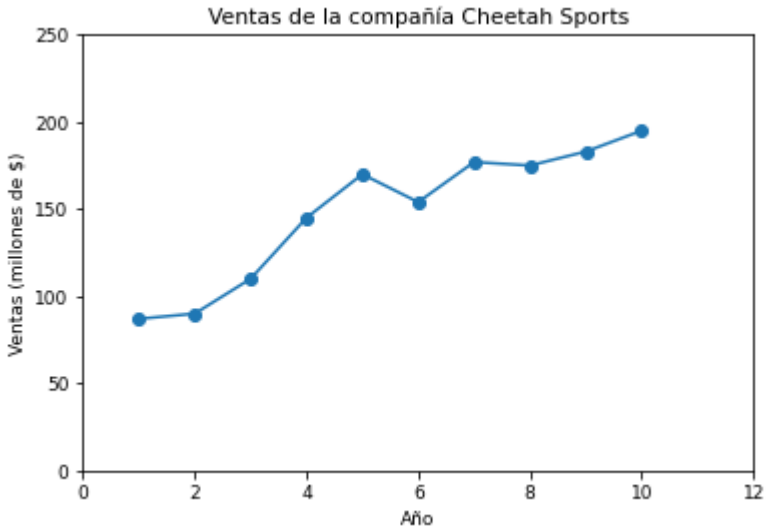
Dataset CheetahRegion

Anio	regionEste	regionOeste	Ventas
1	59	28	87
2	57	33	90
3	68	42	110
4	91	54	145
5	109	61	170
6	96	58	154
7	110	67	177
8	72	103	175
9	63	120	183
10	65	130	195

Los datos de esta tabla ¿Son transversales o corresponden a una serie temporal?

Gráfico de líneas

```
# Genera el grafico de la serie temporal (grafico por defecto)
plt.scatter(data=cheetahRegion, x='Anio', y='Ventas')

# Genera el grafico de la serie temporal
#(mejorando la informacion mostrada)
fig, ax = plt.subplots()

ax.plot('Anio', 'Ventas', data=cheetahRegion, marker="o")

ax.set_title('Ventas de la compañía Cheetah Sports')
ax.set_xlabel('Año', fontsize='medium')
ax.set_ylabel('Ventas (millones de $)', fontsize='medium')
ax.set_xlim(0, 12)
ax.set_ylim(0, 250)
```



Tipo de gráfico muy utilizado para series temporales

Gráfico de líneas

```
# Genera el grafico de ambas series temporales (mejorando la informacion mostrada)
fig, ax = plt.subplots()

# Grafica la serie regionEste
ax.plot('Anio', 'regionEste', data=cheetahRegion,
        marker='.',                    # Tipo de punto (redondo, triángulo, cuadrado, etc.)
        linestyle='-',          # Tipo de linea (solida, punteada, etc.)
        linewidth=0.5,                 # Ancho de linea
        label='Región Este',    # Etiqueta que va a mostrar en la leyenda
        )

# Grafica la serie regionOeste
ax.plot('Anio', 'regionOeste', data=cheetahRegion,
        marker='.',             # Tipo de punto (redondo, triángulo, cuadrado, etc.)
        linestyle='-',          # Tipo de linea (solida, punteada, etc.)
        linewidth=0.5,          # Ancho de linea
        label='Región Oeste'    # Etiqueta que va a mostrar en la leyenda
        )

# Agrega titulo, etiquetas a los ejes y limita el rango de valores de los ejes
ax.set_title('Ventas de la compañía Cheetah Sports según región')
ax.set_xlabel('Año')
ax.set_ylabel('Ventas (millones de $)')
ax.set_xlim(0,12)
ax.set_ylim(0,140)
# Muestra la leyenda
ax.legend()
```

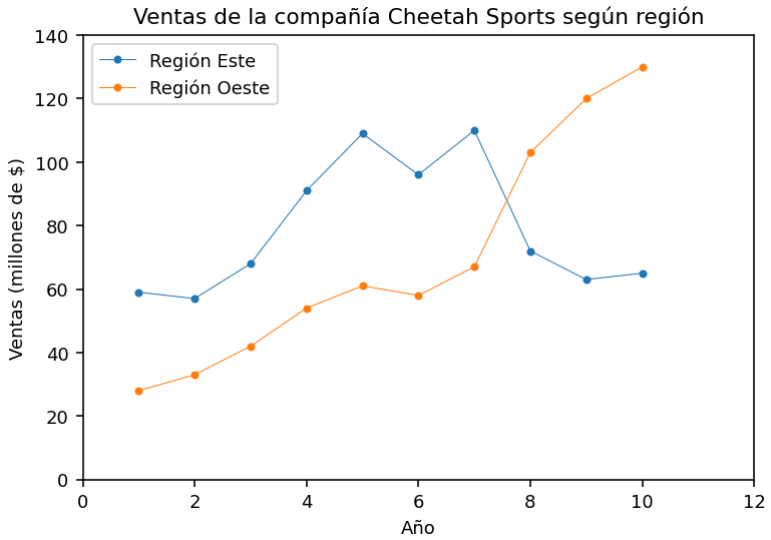


Gráfico de líneas - características

- Representa dos variables numéricas (x,y) donde para cada x hay como mucho un valor de y. La línea graficada interpola linealmente los valores de (x,y) existentes en la muestra.
  - La variable x a veces representa el paso del tiempo.
  - Idealmente debe haber bastantes valores distintos de x para que el gráfico cobre sentido.
- Se pueden representar más variables, mediante el uso de colores, o tipos de línea (llena/punteada/etc).
- Puede ser útil para entender la relación entre las variables.

Ejercicios

Considerar los siguientes datos correspondientes a los precios del biodiesel en distintos períodos en la Argentina
(se encuentran subidos en el campus)

Periodo	Precio
202312	686,986
202311	520
202310	434,006
202309	361,672
202308	346,000
202307	

1. Representarlos gráficamente
2. Analizar los resultados obtenidos
3. Discutir con el resto de la clase
    a. ¿Cuál fue su objetivo: Explorar, Explicar, Otro?
    b. ¿Qué tipos de variables estaban en juego?
    c. ¿Qué tipo de gráfico decidió utilizar?
    d. ¿Qué resultados obtuvo?
    e. ¿Mejoró alguna característica del gráfico para cumplir con el objetivo?

Gráfico de barras

```
#### Genera el grafico de barras de las ventas mensuales
fig, ax = plt.subplots()

ax.bar(data=cheetahRegion, x='Anio', height='Ventas')

ax.set_title('Ventas de la compañía Cheetah Sports')
ax.set_xlabel('Año', fontsize='medium')
ax.set_ylabel('Ventas (millones de $)', fontsize='medium')
ax.set_xlim(0, 11)
ax.set_ylim(0, 250)

ax.set_xticks(range(1,11,1))              # Muestra todos los ticks del eje x
ax.set_yticks([])                         # Remueve los ticks del eje y
ax.bar_label(ax.containers[0], fontsize=8) # Agrega la etiqueta a cada barra
```

¿Por qué podemos usar un BarPlot para este dataset?

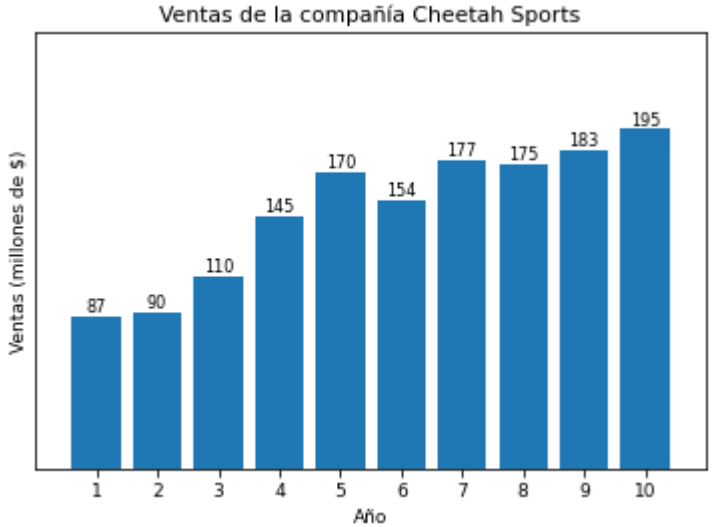


Gráfico de barras agrupadas

```
#### Genera el grafico de barras de ambas series temporales
fig, ax = plt.subplots()

x = cheetahRegion['Anio']
east = cheetahRegion['regionEste']
west = cheetahRegion['regionOeste']

width = 0.4

ax.bar(x - width/2, east, width=width, label='Region Este')
ax.bar(x + width/2, west, width=width, label='Region Oeste')

ax.set_title('Ventas de la compañía Cheetah Sports según región')
ax.set_xlabel('Año')
ax.set_xticks([1,2,3,4,5,6,7,8,9,10])
ax.set_ylabel('Ventas (millones de $)')
ax.legend()

plt.show()
```

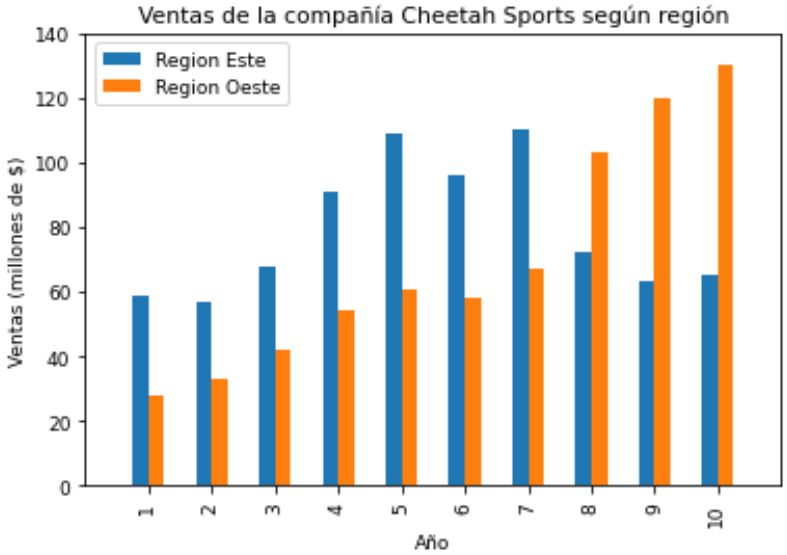


Gráfico de barras apiladas

```
# Genera el grafico de barras apiladas de ambas series temporales
fig, ax = plt.subplots()

# Grafica la serie regionEste
ax.bar(cheetahRegion['Anio'], cheetahRegion['regionEste'] ,
       label='Region Este', color = "#4A4063")
# Grafica la serie regionOeste
ax.bar(cheetahRegion['Anio'], cheetahRegion['regionOeste'],
       bottom=cheetahRegion['regionEste'], label='Region Oeste',
       color = '#BFACC8')

# Agrega titulo, etiquetas a los ejes y limita el rango de valores
de los ejes
ax.set_title('Ventas de la compañía Cheetah Sports según región')
ax.set_xlabel('Año')
ax.set_ylabel('Ventas (millones de $)')
ax.set_xlim(0,10.9)
ax.set_ylim(0,250)
ax.set_xticks(range(1,11,1))    # Muestra todos los ticks del eje x

plt.legend()                    # Muestra la leyenda
```

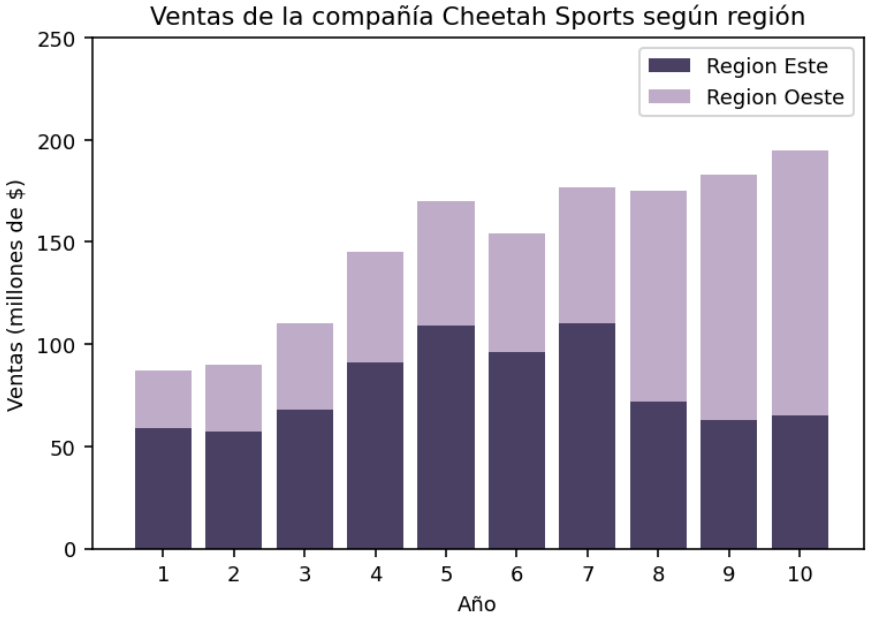


Gráfico de barras - características

- Representa dos variables. Una numérica (altura de las barras) y una categórica u ordinal (una barra para cada categoría).
  - A veces corresponden a una secuencia temporal
  - La altura de las barras suele representar una cantidad, (ej. No debería representar un año)
- Pueden representar además otra variable categórica, mediante el uso de colores.
  - Esta nueva variable debería tomar pocos valores (2 o 3) para que el gráfico no se vuelva incomprensible.
  - Sirve para comparar distribuciones.

Gráfico de torta/*pie chart*

Antes de graficar…

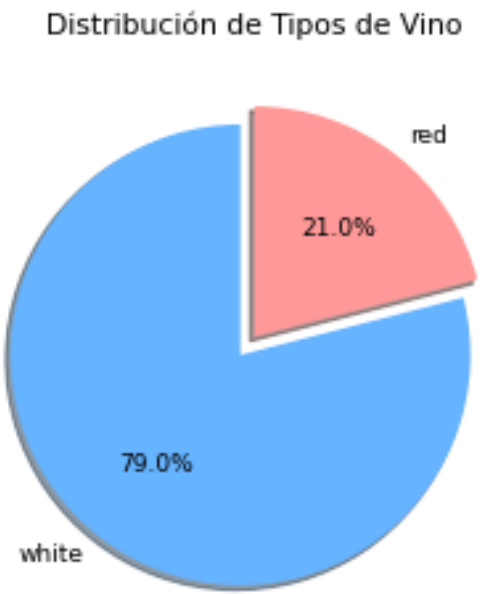
```
# Contamos cuantos vinos de cada tipo hay en el dataset
wine['type'].value_counts()
```

```
Out[5]:
type
white    79
red      21
Name: count, dtype: int64
```

type	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	pH	sulphates	alcohol	quality
white	8.4	0.17	0.31	6.7	0.038	29	132	3.1	0.32	10.6	7
white	6	0.18	0.31	1.4	0.036	14	75	3.34	0.58	11.1	8
white	8.6	0.36	0.26	11.1	0.03	43.5	171	3.03	0.49	12	5
white	6.9	0.4	0.17	12.9	0.033	59	186	3.08	0.49	9.4	5
red	6.8	0.785	0	2.4	0.104	14	30	3.52	0.55	10.7	6
red	10.8	0.29	0.42	1.6	0.084	19	27	3.28	0.73	11.9	6

Gráfico de torta

```
# Transformamos la salida de value_counts enun dataframe
conteos = pd.DataFrame(wine['type'].value_counts()).reset_index()
conteos = conteos.rename(columns={'index': 'type', 0: 'count'})

# Genera el grafico de barras torta (mejorando la informacion mostrada)
fig, ax = plt.subplots()

ax.pie(data=conteos,
       x='count',
       labels='type',           # Etiquetas
       autopct='%1.2f%%',       # porcentajes
       colors=['gold',
               'purple'],
       shadow = True,
       explode = (0.1,0)        # separa las slices del pie plot
       )
```



¿De qué tipo son las variables graficadas?
¿Cualitativas o cuantitativas?

Gráfico de torta

- Un gráfico de torta (*pie plot/pie chart*) es un círculo dividido en porciones que representan partes de un conjunto
- Los humanos **no somos buenos para leer ángulos**

Hagamos un experimento para comprobarlo:

1. Tratemos de identificar al grupo más grande
2. Tratemos de ordenar a los grupos del mayoritario al minoritario

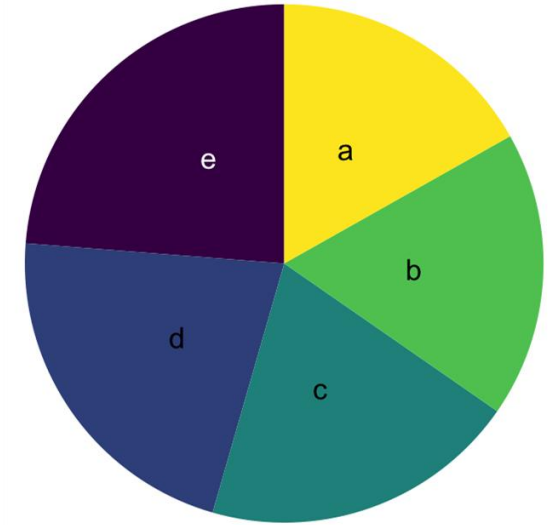


Gráfico de torta

Hagamos un experimento para comprobarlo:

1. Tratemos de identificar al grupo más grande
2. Tratemos de ordenar a los grupos del mayoritario al minoritario

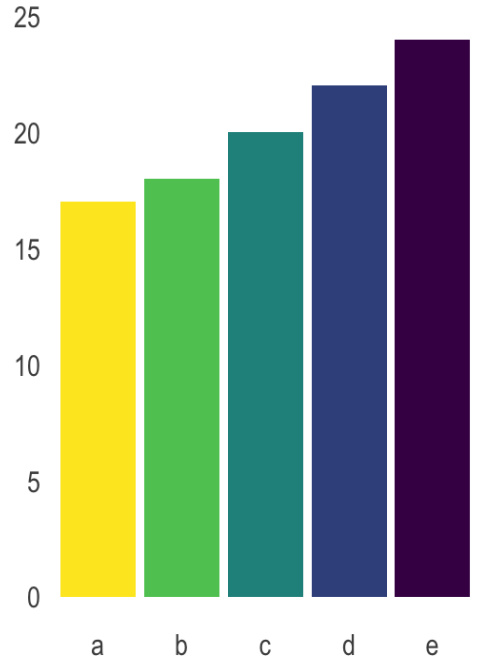
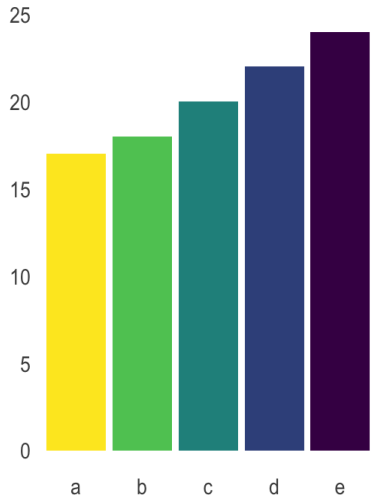


Gráfico de torta

Los dos gráficos fueron construidos usando **los mismos datos,** sin embargo encontrar al grupo mayoritario y ordenar los grupos resulta **mucho más sencillo** observando el **gráfico de barras.**



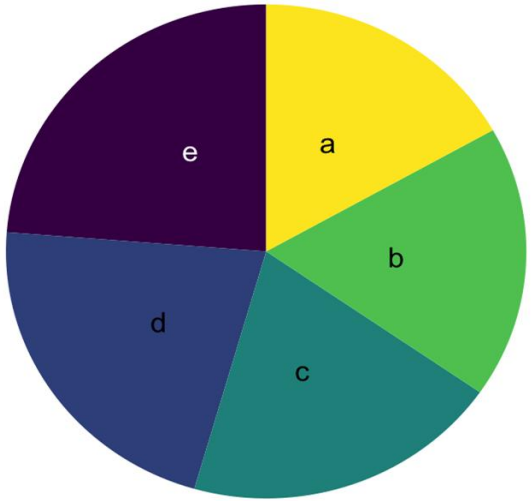


Gráfico de torta - características

- Representa proporciones o distribuciones porcentuales de una variable numérica (en el tamaño de las porciones) respecto de una variable categórica (color)
- Puede ser útil para entender la distribución entre las categorías.
- Puede no ser útil, si hay muchas categorías, o si hay proporciones similares.

Ejercicios

Considerar los siguientes datos a poseedores de teléfonos del archivo telefonosInteligentes.csv

RangoEtario	Telefono_Inteligente (%)	Telefono_NoInteligente (%)	SinTelefono (%)
18-24	49	46	5
25-34	58	35	7
35-44	44	45	11
45-54	28	58	14
55-64	22	59	19
65+	11	45	44

1. Generar un gráfico para representarlos gráficamente
2. Analizar los resultados obtenidos
3. Discutir con el resto de la clase
    a. ¿Cuál fue su objetivo: Explorar, Explicar, Otro?
    b. ¿Qué tipos de variables estaban en juego?
    c. ¿Qué tipo de gráfico decidió utilizar?
    d. ¿Qué resultados obtuvo?
    e. ¿Mejoró alguna característica del gráfico para cumplir con el objetivo?
    f. Responder Verdadero o Falso y justificar visualmente. **“Es más probable que las personas mayores posean un teléfono inteligente a que las personas más jóvenes posean uno inteligente.”**

Buenas Prácticas

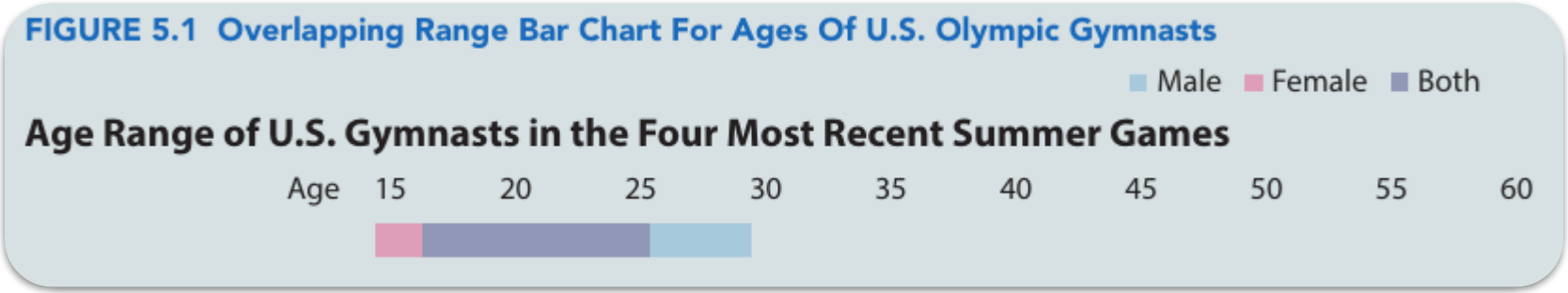
- Elegir el tipo de gráfico adecuado
- Usar colores con sentido
- Usar pocos colores y diferenciables
- Hacer gráficos que aporten información útil
- No agregar mucha información en un solo gráfico
- Priorizar legibilidad frente a estética

Distribución de los Datos

Visualización - Distribución de los Datos

Edades de los atletas olímpicos de USA en los últimos 4 torneos de verano

(Bordes: izquierdo -> mínimo; derecho -> máximo)

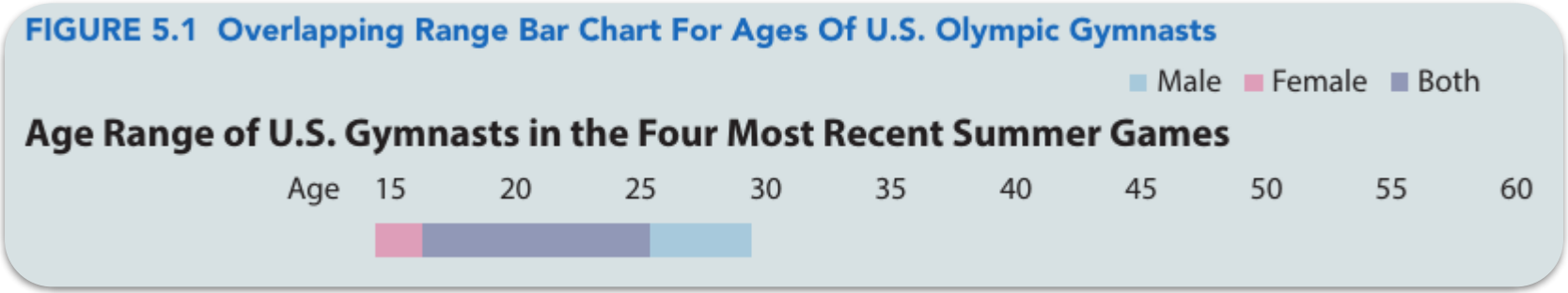


1. ¿En qué rango de edades hubo participación femenina? ¿Masculina? ¿Ambos?
2. ¿Cuantos individuos femeninos de 20 años de edad participaron? ¿Y masculinos?
3. ¿Para qué edades podemos afirmar que hubo participación?

Visualización - Distribución de los Datos

Edades de los atletas olímpicos de USA en los últimos 4 torneos de verano

(Bordes: izquierdo -> mínimo; derecho -> máximo)



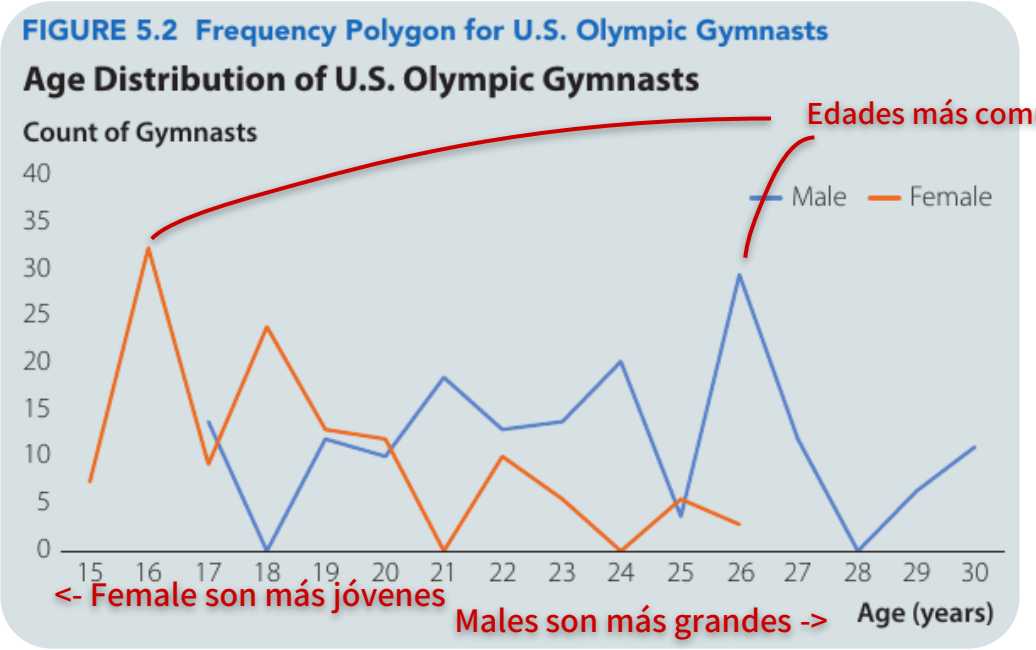
¡No hay información sobre cómo se distribuyen lxs gimnastxs femeninos y masculinos en sus respectivos rangos!

Esta forma de visualizar datos:

- No es intuitiva
- Aumenta la carga cognitiva de la audiencia

Visualización - Distribución de los Datos

Edades de los atletas olímpicos de USA en los últimos 4 torneos de verano



1. ¿En qué rango de edad hubo participación femenina?¿masculina?¿ambos?
2. ¿Qué cantidad de individuos de sexo femenino, de 20 años, participó?¿y de sexo masculino?¿y de ambos?
3. ¿Para qué edades podemos afirmar que hubo participación?

¡Muestra información sobre la distribución por edades de gimnastxs masculinos y femeninos!

Visualización - Distribución de los Datos

El rol del **análisis descriptivo** es analizar y visualizar datos para comprender mejor la variación y su impacto

Distribución de frecuencia de una variable:
Describe qué valores se observaron y con qué frecuencia esos valores aparecen en dichos datos

Distribución de frecuencia

- **Variable categórica**
  (etiquetas que no pueden manipularse aritméticamente)
- **Variable cuantitativa**
  (valores numéricos que pueden manipularse aritméticamente)

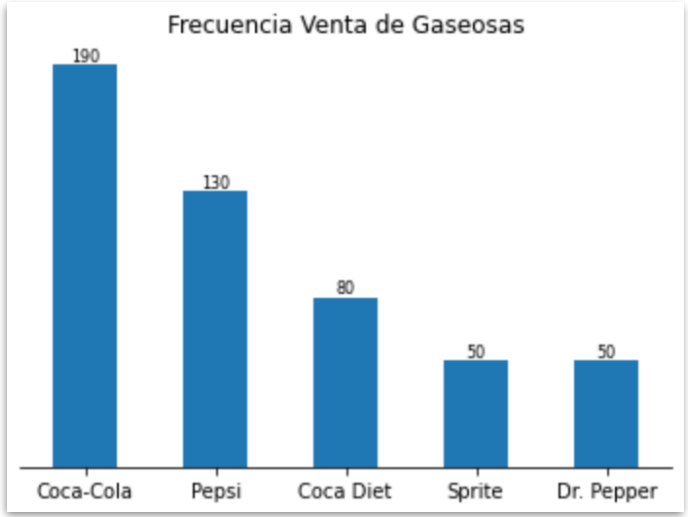
Visualización - Distribución de los Datos - Categóricos

Una distribución de frecuencia (de variable categóricas) es un resumen de datos que muestra el **número** (frecuencia) de observaciones en **cada una de las clases** (no superpuestas), denominadas **bins**.

Compras_gaseosas
Coca-Cola
Sprite
Pepsi
Pepsi
Dr. Pepper
Coca Diet
Coca-Co

500 compras

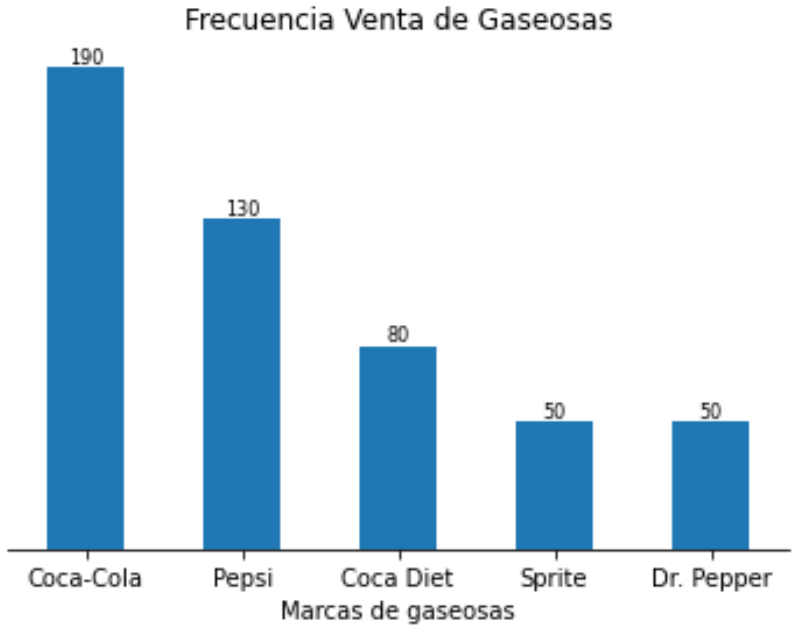
Compras_gaseosas	
Coca-Cola	190
Pepsi	130
Coca Diet	80
Dr. Pepper	50
Sprite	50



La distribución de frecuencia resume la información sobre la popularidad de las cinco gaseosas

Gráfico - Distribución de Datos categóricos

```
fig, ax = plt.subplots()
gaseosas['Compras_gaseosas'].value_counts().plot.bar(ax = ax)


ax.set_title('Frecuencia Venta de Gaseosas')
ax.set_xlabel('Marcas de gaseosas')
ax.set_yticks([])                              # Remueve los ticks del eje y
ax.bar_label(ax.containers[0], fontsize=8)     # Agrega la etiqueta a cada barra
ax.tick_params(axis='x', labelrotation=0)      # Rota las etiquetas del eje x

# Eliminar lineas del recuadro
ax.spines[['right', 'top', 'left']].set_visible(False)
```



Distribución de Datos categóricos

Distribución de **Frecuencia Absoluta** muestra la **cantidad** (recuento) de artículos en cada uno de los bins. A veces nos interesa la **proporción o porcentaje** de artículos en cada contenedor.

Frecuencia relativa de un bin. Fracción o proporción de ítems que pertenecen a ese bin (clase).

$$\text{Frecuencia relativa de un bin} = \frac{\text{Frecuencia absoluta del bin}}{n}$$

donde n es la cantidad total de observaciones

Frecuencia porcentual de un bin. Frecuencia relativa multiplicada por 100.

Visualización - Distribución de los Datos - Categóricos

Compras_gaseosas
Coca-Cola
Sprite
Pepsi
Pepsi
Dr. Pepper
Coca Diet
Coca-Co

Frecuencia relativa

Compras_gaseosas	
Coca-Cola	0.38
Pepsi	0.26
Coca Diet	0.16
Dr. Pepper	0.10
Sprite	0.10

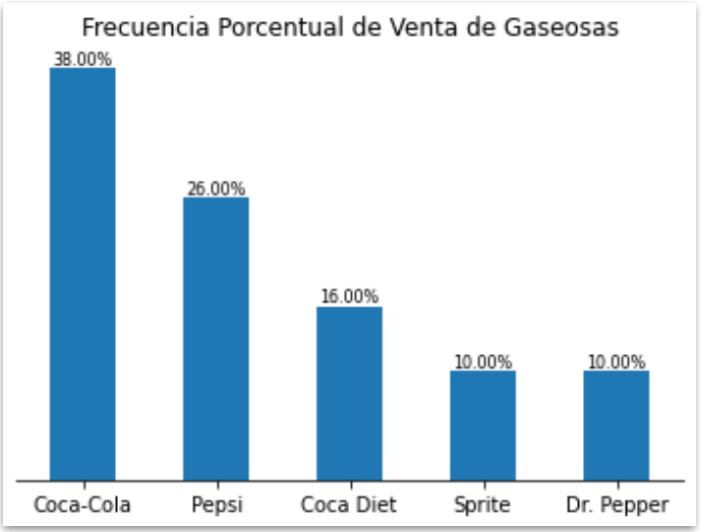
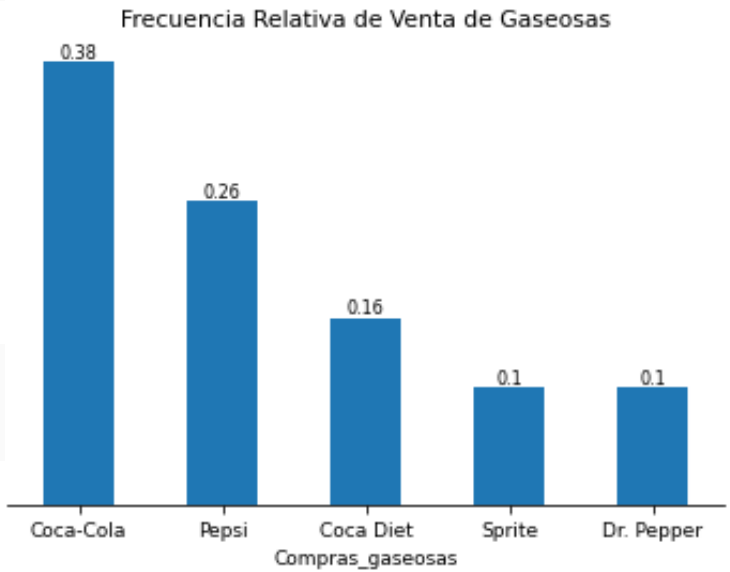


Gráfico - Distribución de Datos categóricos

```
# Tabla de frecuencias relativas
gaseosas['Compras_gaseosas'].value_counts(normalize=True)
```

```
Out[138]:
Compras_gaseosas
Coca-Cola     0.38
Pepsi         0.26
Coca Diet     0.16
Sprite        0.10
Dr. Pepper    0.10
Name: proportion, dtype: float64
```

```
fig, ax = plt.subplots()

ax = gaseosas['Compras_gaseosas'].value_counts(normalize=True).plot.bar()
```



Distribución de Datos categóricos

- Distribución de frecuencia relativa (o porcentual) se puede usar para estimar las probabilidades relativas de diferentes valores para una variable (aleatoria)

- Ej. Un puesto de comida ha determinado que adquirirá un total de 12.000 gaseosas para un próximo concierto. ¿Cómo dividirían este total entre los distintos tipos de gaseosas individuales?

  Si los datos analizados (muestra) son representativos de la población de clientes del puesto de comida, se puede usar esta información para determinar los volúmenes apropiados de cada tipo de refresco.
  Por ejemplo, los datos sugieren que se debería adquirir 12.000 * 0,38 = 4.560 Coca-Colas.

Compras_gaseosas	
Coca-Cola	0.38
Pepsi	0.26
Coca Diet	0.16
Dr. Pepper	0.10
Sprite	0.10

Distribución de Datos continuos

¿Cuál es la dificultad para obtener la distribución de frecuencia en variables continuas?
Por ejemplo la distribución para la variable **Peso** (continua, asumir 3 decimales)

Peso (kg)
59,032
78,127
95,900

Distribución de Datos continuos

¿Cuál es la dificultad para obtener la distribución de frecuencia en variables continuas?
Por ejemplo la distribución para la variable **Peso** (continua, asumir 3 decimales)

Peso (kg)
59,032
78,127
95,900

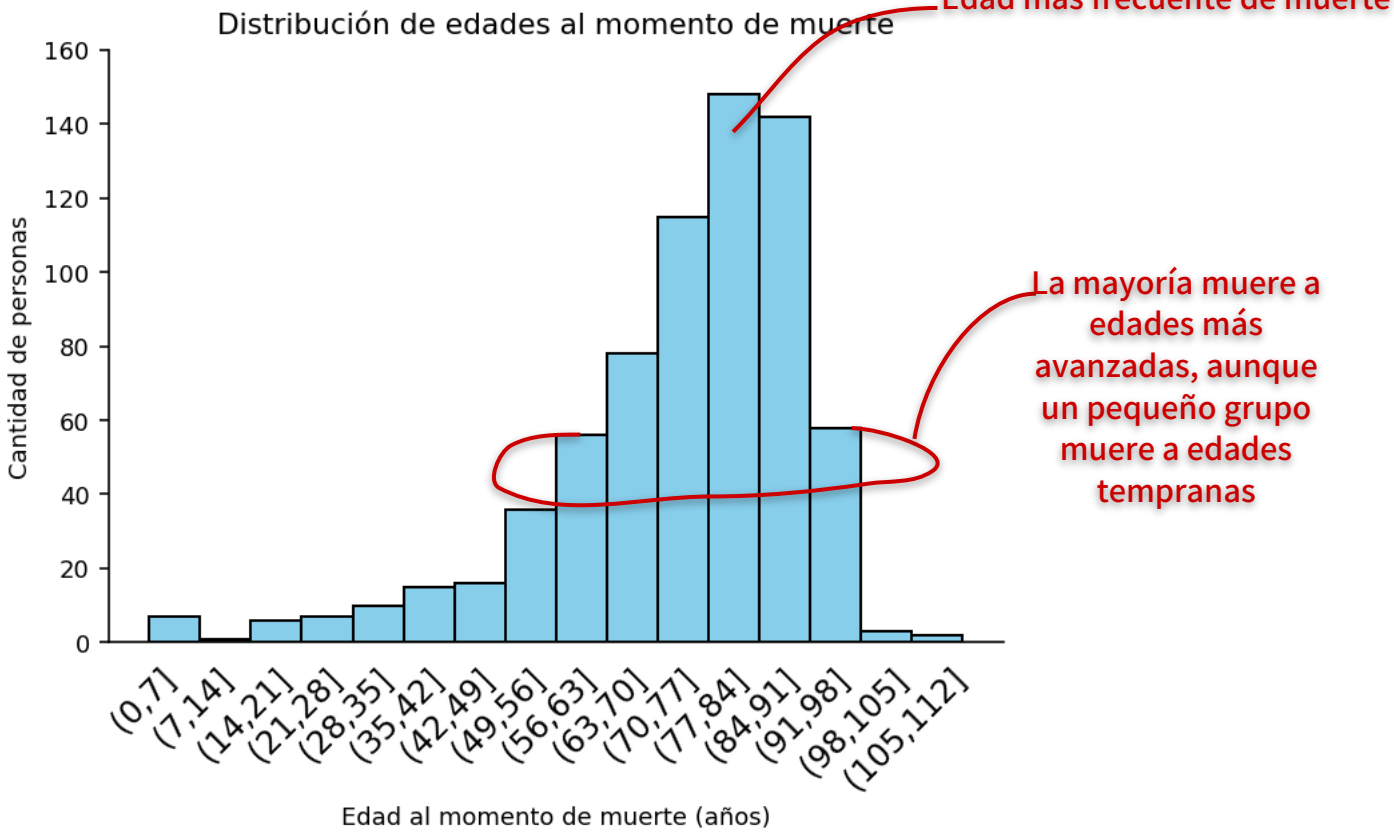
Solución:

- Cada bin ahora contiene un **rango de valores** (en vez de 1 solo valor)
- Como antes, los bins **no se deben superponer**

Distribución de Datos continuos

AgeAtDeath
81
64
88
85
96
101
87

700 datos

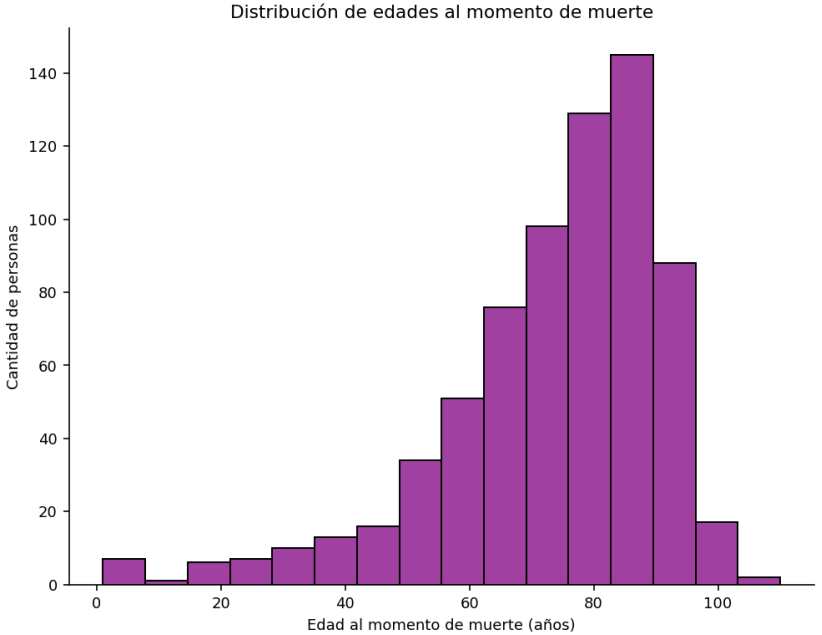


Edad más frecuente de muerte

La mayoría muere a edades más avanzadas, aunque un pequeño grupo muere a edades tempranas

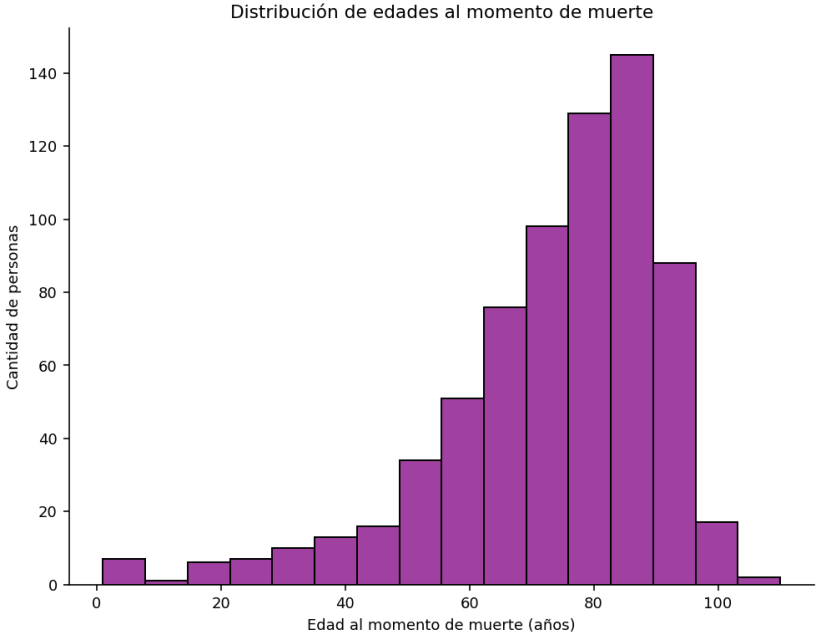
Distribución de Datos continuos

```
import seaborn as sns
#Graficar la distribucion usando seaborn
plt.figure(figsize=(8, 6))
sns.histplot(ageAtDeath['AgeAtDeath'], kde=False, bins=16,
color='purple') # kde=True agrega la curva de densidad

plt.title('Distribución de edades al momento de muerte')
plt.xlabel('Edad al momento de muerte (años)')
plt.ylabel('Cantidad de personas')

# Mostrar el grafico
plt.show()
```



Distribución de Datos continuos

```
import seaborn as sns
#Graficar la distribucion usando seaborn
plt.figure(figsize=(8, 6))
sns.histplot(ageAtDeath['AgeAtDeath'], kde=False, bins=16,
color='purple') # kde=True agrega la curva de densidad

plt.title('Distribución de edades al momento de muerte')
plt.xlabel('Edad al momento de muerte (años)')
plt.ylabel('Cantidad de personas')

# Mostrar el grafico
plt.show()
```



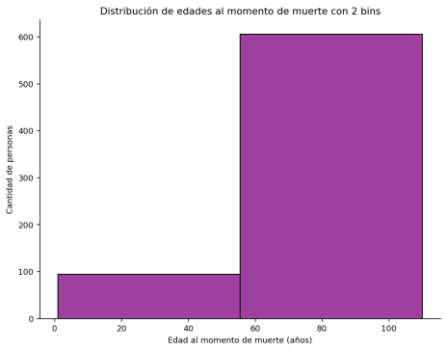
¿Qué pasa si cambiamos la cantidad de bins? Grafiquen usando los valores comentados en el código

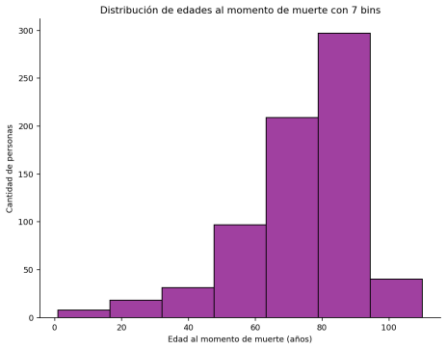
Distribución de Datos continuos

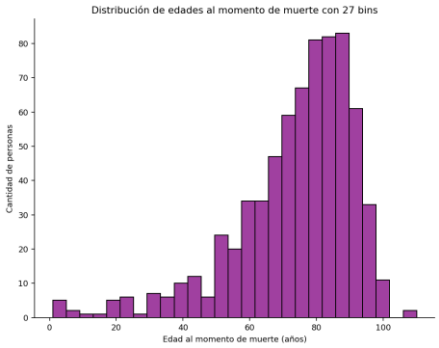
La cantidad de bins y el ancho de los mismos puede afectar en gran medida la visualización de una distribución

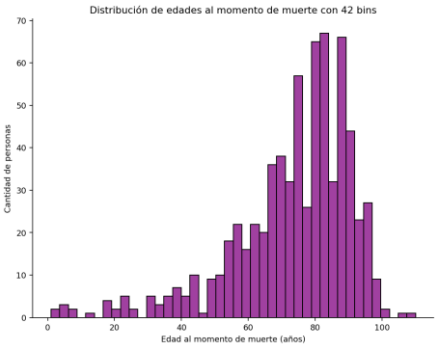
Para graficar tenemos que definir:

- La cantidad de bins
- El ancho (rango numérico) de cada bin
- El rango total que abarca el conjunto de bins









Distribución de Datos continuos

1. **Cantidad de bins**

   Muchos bins → contienen sólo unas pocas observaciones → no captura patrones generalizables (puede parecer irregular y "ruidoso")

   Pocos bins → rango de valores muy amplio en mismo bin → no captura con precisión la variación en los datos y solo presenta patrones “borrosos” de alto nivel.

La elección de la cantidad de bins es **subjetiva**, depende del tema y el objetivo del análisis.

Recomendación: utilizar de 5 a 20 bins

Pocas observaciones → 5 o 6 bins

Muchas observaciones → Más bins.

Distribución de Datos continuos

2. Ancho de bins

Tomar anchos distintos para cada bin puede llevar a decisiones equivocadas

Recomendación: utilizar bins del mismo ancho

Distribución de Datos continuos

2. Rango de valores de bins

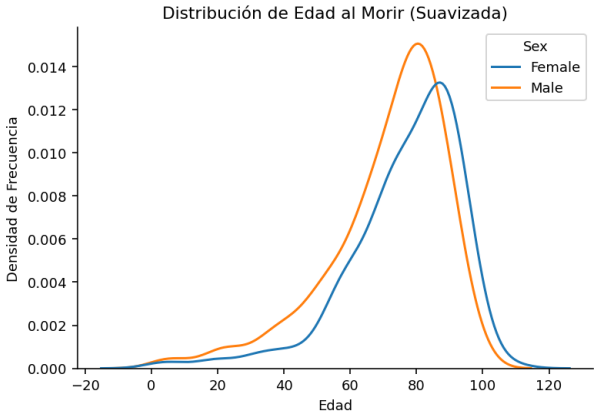
Todas las observaciones deberían caer dentro de un bin
Los bins **no deben superponerse**
Tener cuidado con los extremos

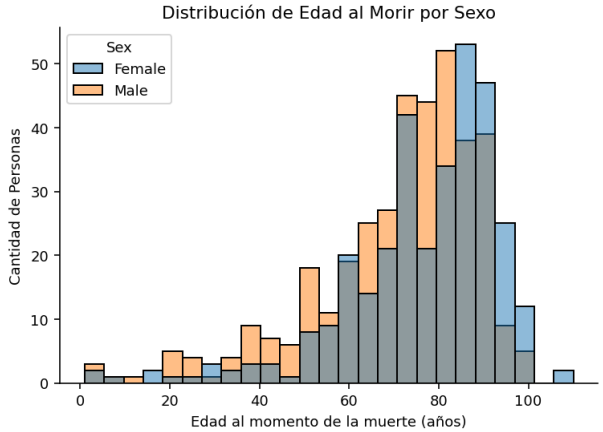
Recomendación: utilizar rangos de bins que cumplan con lo anterior

Distribución de Datos continuos

¿Y si queremos analizar la variabilidad de dos variables?

Sex	AgeAtDeath
Female	81
Female	64
Male	88
Female	85
Female	96
Female	101
Female	87
Male	





Creemos nuestros gráficos

Ejercicio

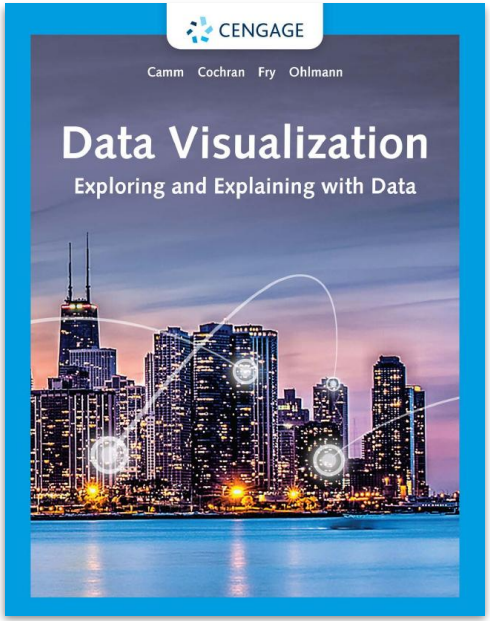
Consigna.
Sean los datos correspondientes a las propinas de un bar (están cargados en el campus en el archivo *tips.csv*)

1. Generar un gráfico para analizar la distribución de la propina en función del:
    - Sexo
    - Día de la semana
1. Comentar los resultados obtenidos

Tarea

- Resolver la guía de ejercicios de visualización (hasta ejercicio 12)

Bibliografía



Camm/Cochran/Fry/Ohlmann, Data Visualization: Exploring and Explaining with Data, 1st. Edition, Cengage Learning, 2022