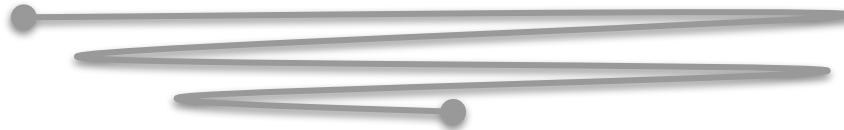


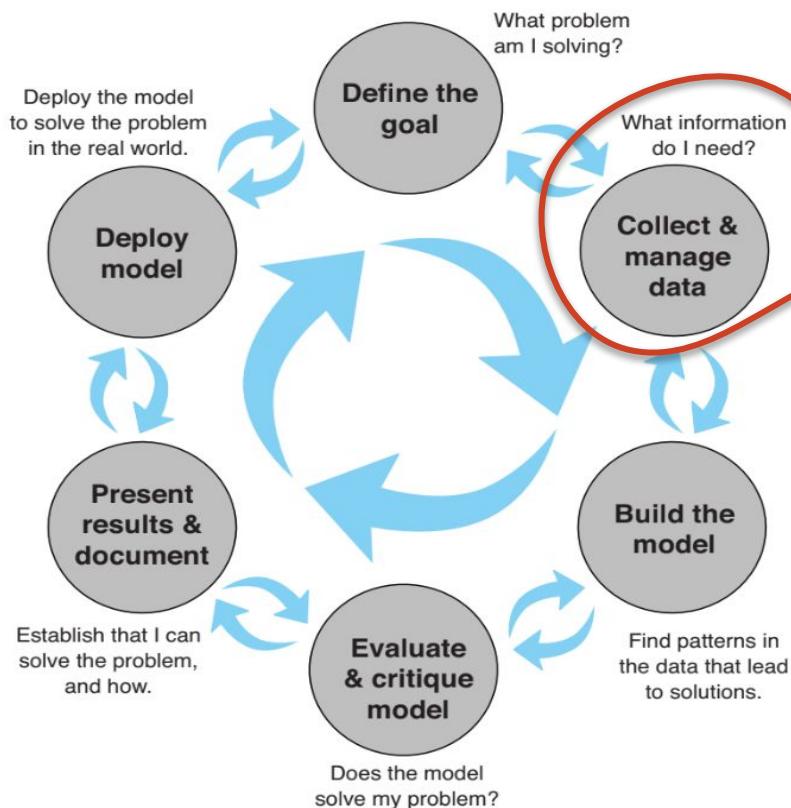
Laboratorio de Datos



Visualización y Análisis Exploratorio de Datos - Parte 01



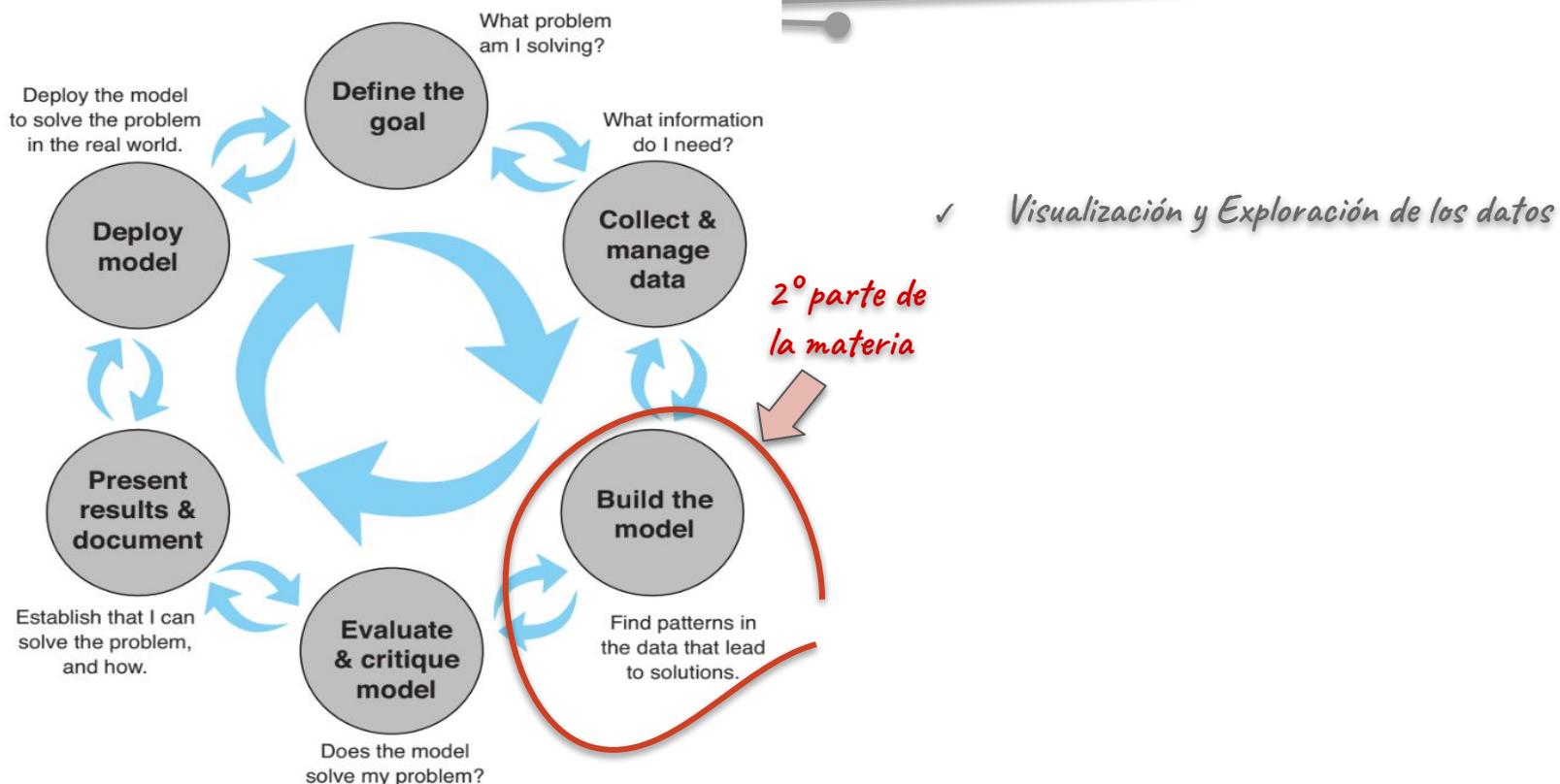
Recorrido de la materia (hasta ahora)



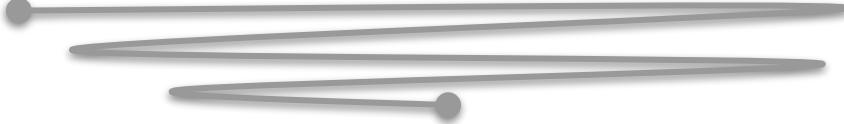
1º parte de
la materia

- ✓ 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
- ✓ Leyes acerca de la Protección de Datos

Recorrido de la materia (clase de hoy)



Preguntas ...



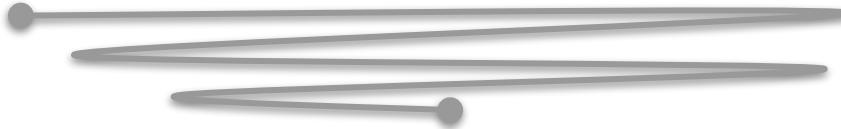
¿Qué es el proceso de análisis de datos? (1 min ...)

El análisis de datos es el proceso científico de transformar datos en información para tomar mejores decisiones

¿Cuáles son los hechos que impulsaron el desarrollo del análisis de datos?

- La generación masiva de datos (sensores, e-commerce, etc.)
- El desarrollo de metodologías (machine learning, optimización, simulación, etc.)
- Los avances en la capacidad computacional y almacenamiento (hardware, parallel computing, cloud computing, etc.)
- ...

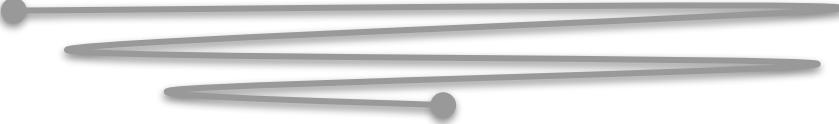
Alcance del Análisis de Datos



- No siempre es posible recopilar datos de toda la **población** (population)
- En esos casos recopilamos datos de un subconjunto de la población: **muestra** (sample)

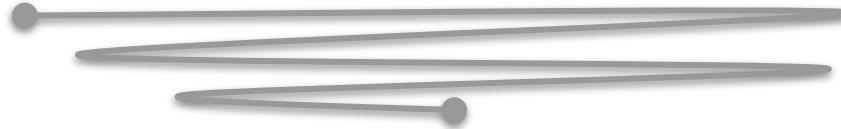
Si asumimos que estamos tratando con una muestra de datos representativa de la población, podemos hacer generalizaciones sobre toda la población

Alcance del Análisis de Datos



- Análisis de Datos. Puede abarcar desde simples reportes hasta modelos y simulaciones
- Según el alcance (metodología) ...
 - **Análisis descriptivo.** Conjunto de herramientas analíticas que describen lo que ha sucedido.
Ej. Queries, reportes, estadística descriptiva, visualización de datos. En general, estas técnicas 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, datamining predictivo.
 - **Análisis prescriptivo.** Son modelos matemáticos o lógicos que sugieren una decisión o un curso de acción. Ej. modelos de optimización matemática, análisis de decisiones y sistemas heurísticos o basados en reglas.

Alcance del Análisis de Datos



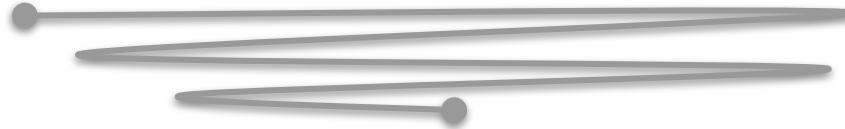
Análisis descriptivo

Análisis predictivo

Análisis prescriptivo

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

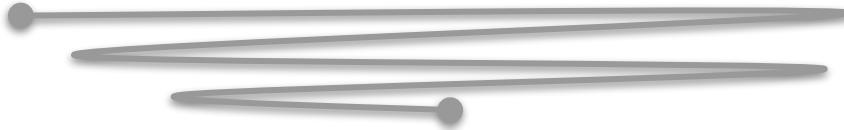
Exploración y Explicación



Existen distintos tipos de **visualización** de los datos según la finalidad

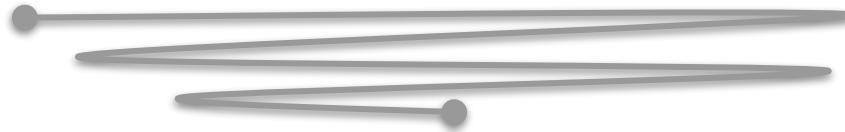
1. Explorar (los datos)
2. Explicar (comunicar/transmitir un mensaje)

Visualización - Exploración y Explicación



1. Exploración de datos

Visualización - Exploración de datos



Ejemplo ¿Cuál es el patrón de asistencia del público al zoológico?

TABLE 1.1

Zoo Attendance Data

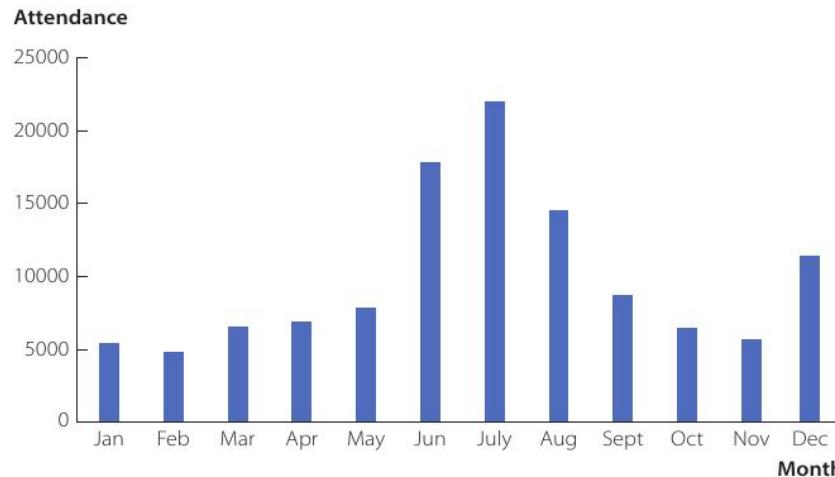
Month	Jan	Feb	Mar	Apr	May	Jun
Attendance	5422	4878	6586	6943	7876	17843
Month	July	Aug	Sept	Oct	Nov	Dec
Attendance	21967	14542	8751	6454	5677	11422

Visualización - Exploración de datos

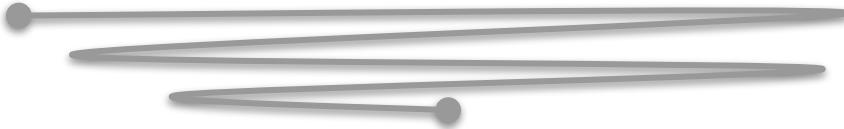
Ejemplo ¿Cuál es el patrón de asistencia del público al zoológico?

FIGURE 1.1

A Column Chart of Zoo Attendance by Month



Visualización - Exploración de datos



Para poner en contexto ...

Se sabe que la asistencia al zoo es mayor en los meses de verano, cuando los niños en edad escolar no asisten a la escuela (vacaciones de verano). Además, la asistencia aumenta gradualmente a partir del mes de febrero hasta mayo a medida que aumenta la temperatura promedio, y la asistencia disminuye gradualmente a partir del mes de septiembre hasta noviembre a medida que disminuye la temperatura promedio.

Pero ... ¿por qué la asistencia al zoo en diciembre y enero no sigue estos patrones (hay una irregularidad)?

El zoo cuenta con el “Festival de las Luces” que se extiende desde finales de noviembre hasta principios de enero. Los niños no van a la escuela durante la última quincena de diciembre y principios de enero debido a las vacaciones, lo que provoca una mayor asistencia por las tardes al zoológico a pesar de las temperaturas intermedias más frías.

La exploración visual de datos es una parte importante del análisis descriptivo

Visualización - Exploración de datos

La exploración de datos permite ...

- Identificar patrones
- Reconocer anomalías o irregularidades
- Conocer mejor la relación entre variables

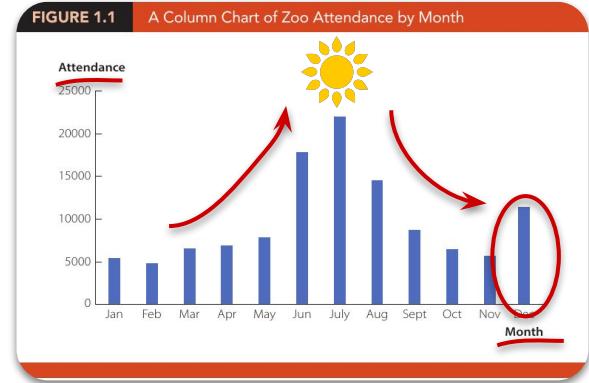
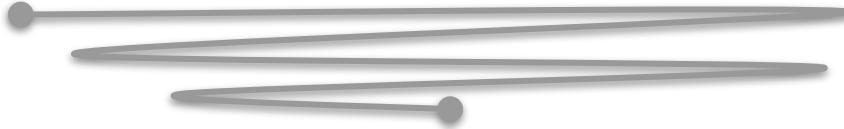


TABLE 1.1 Zoo Attendance Data

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

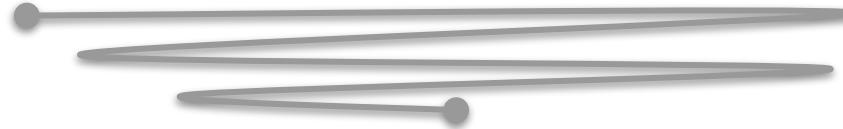
La visualización nos da mayor capacidad para detectar patrones, anomalías y relaciones entre variables que al hacerlo mirando simplemente los datos crudos

Visualización - Exploración y Explicación



2. Explicación

Visualización - Explicación

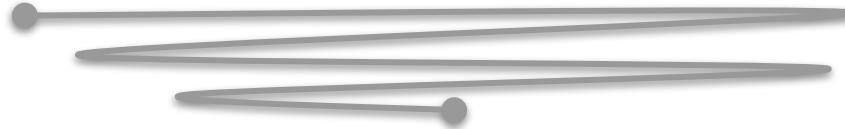


Ejemplo. La “cultura de la compañía” se encuentra entre los factores más importantes a la hora de buscar un trabajo

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%

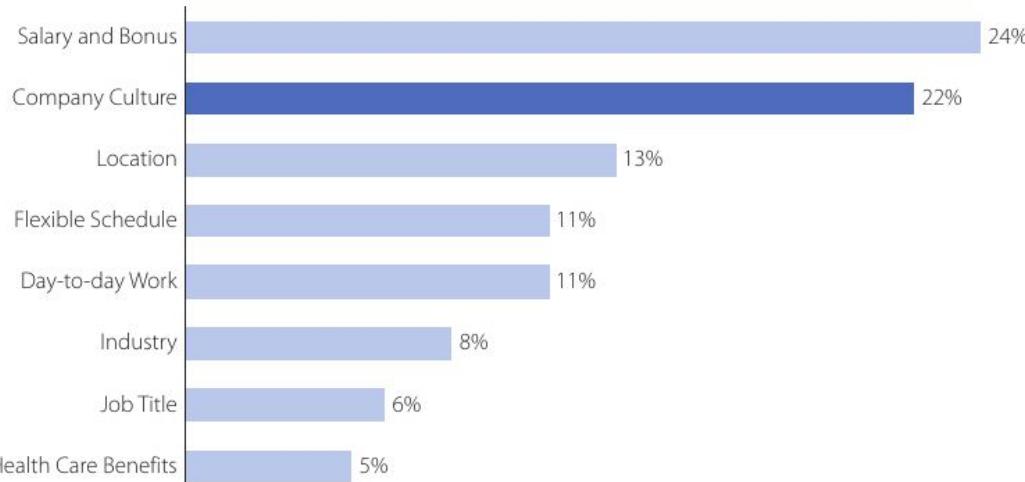
¿Resulta obvio?

Visualización - Explicación



Ejemplo. La “**cultura de la compañía**” se encuentra entre los factores más importantes a la hora de buscar un trabajo

What matters most to you when deciding which job to take next?



¿Resulta obvio?

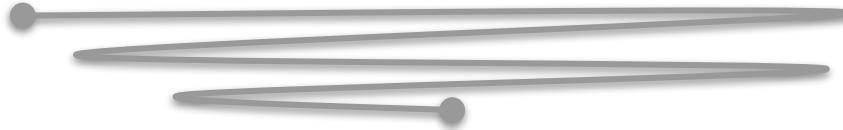
La visualización de datos es útil para comunicar a la audiencia y garantizar que comprenda y se concentre en el mensaje deseado.

Visualización



Tipos de Datos

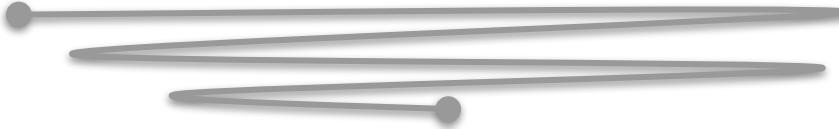
Visualización - Tipos de Datos



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

1. *Cuantitativos vs. Categóricos*
2. *Transversales vs. Series Temporales*
3. *Big Data*

Visualización - Tipos de Datos



1. Cuantitativos vs. Categóricos

Visualización - Cuantitativos vs. Categóricos

¿la variable Volume es cuantitativa o categórica?

Variable cuantitativa

Data for the Dow Jones Industrial Index Companies (April 3, 2020)				
Company	Symbol	Industry	Share Price (\$)	Volume
Apple Inc.	AAPL	Technology	241.41	32,470,017
American Express	AXP	Financial Services	73.6	9,902,194
Boeing	BA	Manufacturing	124.52	36,489,379
Caterpillar Inc.	CAT	Manufacturing	114.67	4,803,174
Cisco Systems	CSCO	Technology	39.06	21,235,157
Chevron	CVX	Petroleum	75.11	14,317,998
Disney	DIS	Entertainment	93.88	14,592,062
Goldman Sachs	GS	Financial Services	146.93	2,773,298
Home Depot, Inc.	HD	Retailing	178.7	6,762,357
IBM	IBM	Technology	106.34	3,909,196
Intel Corporation	INTC	Technology	54.13	23,904,062
Johnson & Johnson	JNJ	Pharmaceutical		
JPMorgan Chase	JPM	Financial S		
	KO			

Permiten indicar una magnitud.

Se les puede aplicar operaciones aritméticas
 $(+, -, \times, \%, \text{ etc.})$.

Ej. podemos sumar los valores de Volumen para calcular el volumen total de todas las acciones negociadas por empresas.

Visualización - Cuantitativos vs. Categóricos

¿La variable Industry es cuantitativa o categórica?

Company	Symbol	Industry	Share Price (\$)	Volume
Apple Inc.	AAPL	Technology	241.41	32,470,017
American Express	AXP	Financial Services	73.6	9,902,194
Boeing	BA	Manufacturing	124.52	36,489,379
Caterpillar Inc.	CAT	Manufacturing	114.67	4,803,174
Cisco Systems	CSCO	Technology	39.06	21,235,157
Chevron	CVX	Petroleum	75.11	14,317,998
Disney	DIS	Entertainment	93.88	14,592,062
Goldman Sachs	GS	Financial Services	146.93	2,773,298
Home Depot, Inc.	HD	Retailing	178.7	6,762,357
IBM	IBM	Technology	106.34	3,909,196
Intel Corporation	INTC	Technology	54.13	23,904,062
Johnson & Johnson	JNJ	Pharmaceutical		
JPMorgan Chase	JPM	Financial S		
	KO			

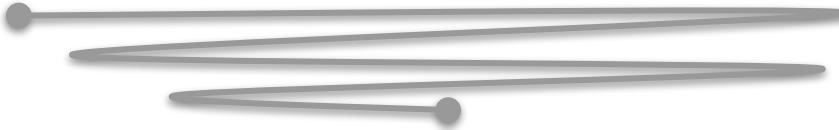
Variable categórica

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.

Ej. Podemos contar el número de empresas que están en la industria tecnológica.

Visualización - Tipos de Datos



2. Transversales vs. Series Temporales

Visualización - Transversales vs. Series Temporales

Los datos de esta tabla ¿son transversales o son series temporales?

TABLE 1.3 Data for the Dow Jones Industrial Index Companies
(April 3, 2020)

Company	Symbol	Industry	Share Price (\$)	Volume
Apple Inc.	AAPL	Technology	241.41	32,470,017
American Express	AXP	Financial Services	73.6	9,902,194
Boeing	BA	Manufacturing	124.52	36,489,379
Caterpillar Inc.	CAT	Manufacturing	114.67	4,803,174
Cisco Systems	CSCO	Technology	39.06	21,235,157
Chevron	CVX	Petroleum	75.11	14,317,998
Disney	DIS	Entertainment	93.88	14,592,062
Goldman Sachs	GS	Financial Services	146.93	2,773,298
Home Depot, Inc.	HD	Retailing	178.7	6,762,357
IBM	IBM	Technology	106.34	3,909,196
Intel Corporation	INTC	Technology	54.13	23,904,062
Johnson & Johnson	JNJ	Pharmaceutical		
JPMorgan Chase	JPM	Financial S		
	KO			

Datos transversales (Cross-Sectional Data)

Son datos que corresponden al mismo momento o aprox. del mismo tiempo.

Ej. Los datos de la tabla son datos transversales porque describen las 30 empresas que componen el índice Dow tomados todos en el mismo momento (abril de 2020).

Visualización - Transversales vs. Series Temporales

Los datos de esta figura, ¿son transversales o son series temporales?

FIGURE 1.4

Dow Jones Index Values from January 2010 to April 2020



datos a través del tiempo

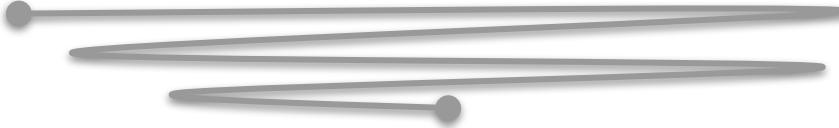
Serie de tiempo (Time Series Data).

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 - Tipos de Datos



3. Big Data

Visualización - Big Data

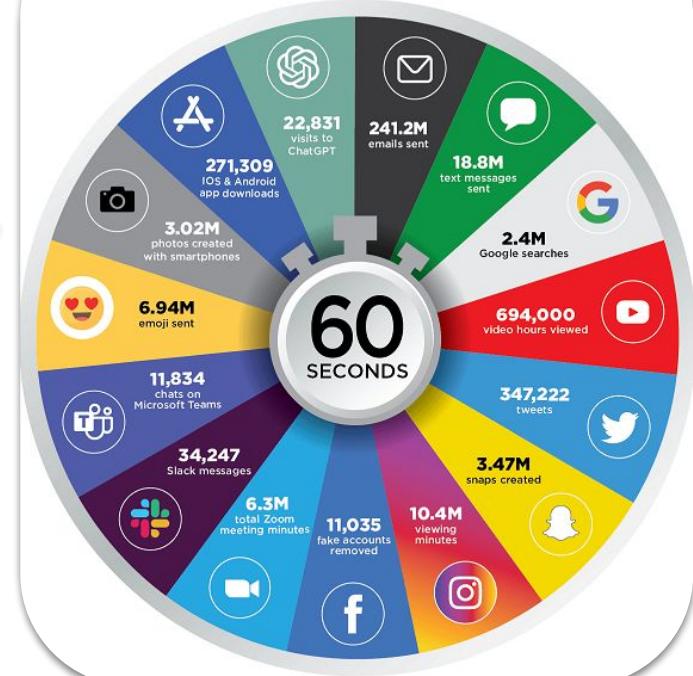
¿Qué es Big Data?

No hay consenso en la definición

Possible definición. Cualquier conjunto de datos que es demasiado grande o demasiado complejo para ser manejado mediante técnicas de procesamiento de datos estándar utilizando una computadora típica de escritorio. Se suele referir a Big Data utilizando las cuatro V:

- *Volumen: la cantidad de datos generados*
- *Velocidad: la velocidad a la que se generan los datos*
- *Variedad: la diversidad de tipos y estructuras de datos generados*
- *Veracidad: la confiabilidad de los datos generados*

THE INTERNET IN 2023 EVERY MINUTE



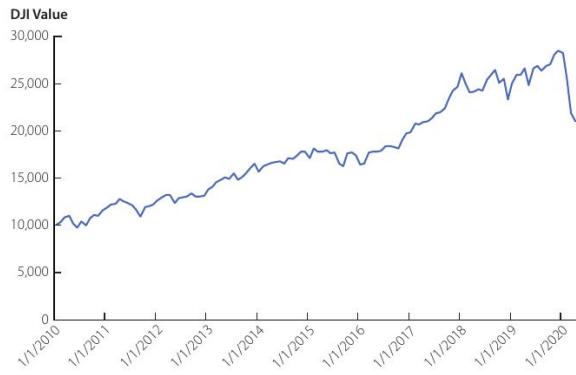
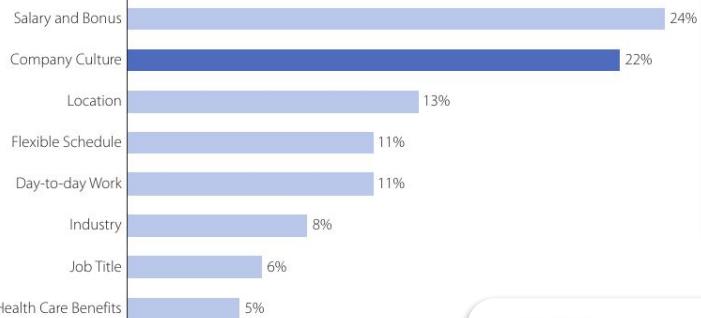
Visualización



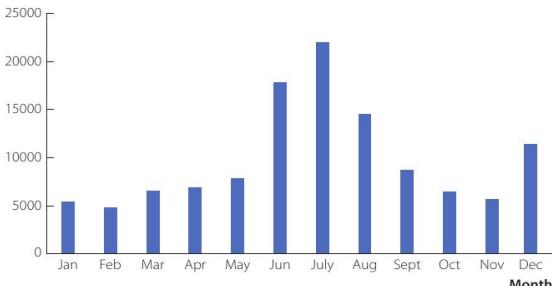
Algunos ejemplos

Visualización - Ejemplos comunes

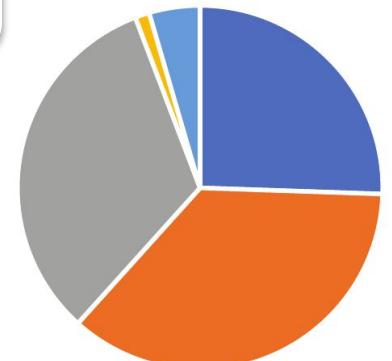
What matters most to you when deciding which job to take next?



Attendance



Number of Vehicles



Fiat Chrysler Ram

GM Chevy Silverado/GMC Sierra

Toyota Tundra

Ford F-Series

Nissa Titan

Pie Chart

Visualización - Ejemplos menos comunes

FIGURE 1.12

A Spaghetti Chart of Hurricane Paths from Multiple Predictive Models

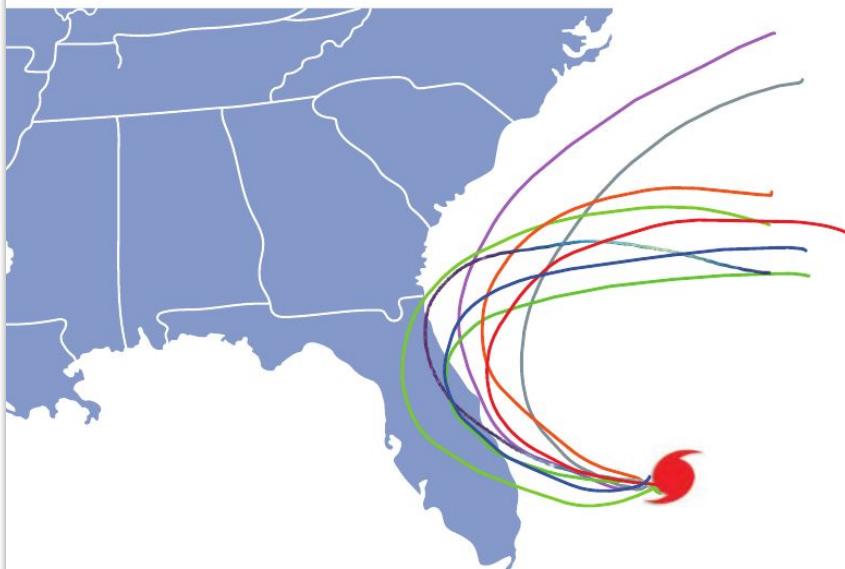
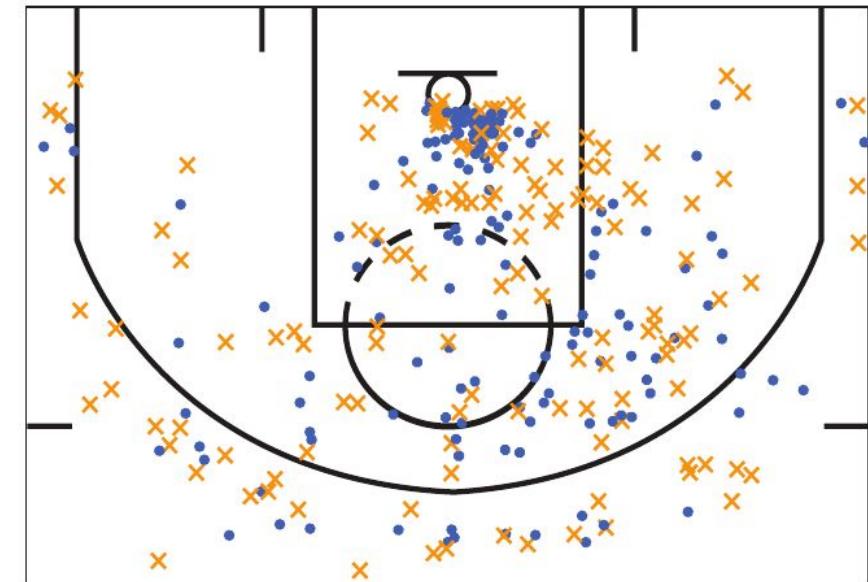


FIGURE 1.13

A Shot Chart for NBA Player Chris Paul



Selección del tipo de gráfico

A la hora de seleccionar un tipo de gráfico hay que tener en cuenta ...

1. Si el objetivo es ...

- 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

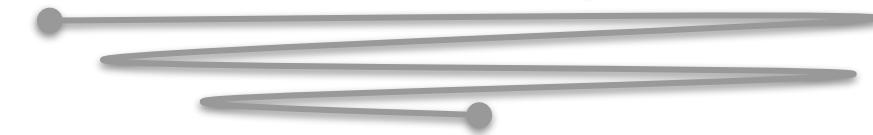
2. El Tipo de datos

- a. Variables **cuantitativas/cualitativas**
- b. Datos **Transversales vs. Series Temporales**
- c. **Big Data**

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 en un día
- d. **Composición.** Entender cómo se constituye una cierta entidad. Ej. Voto de las últimas elecciones

Creemos nuestros gráficos



You can use text.
plt.text(x, y, text)
text is the text you want to print.
x and y are the position of
your text will be

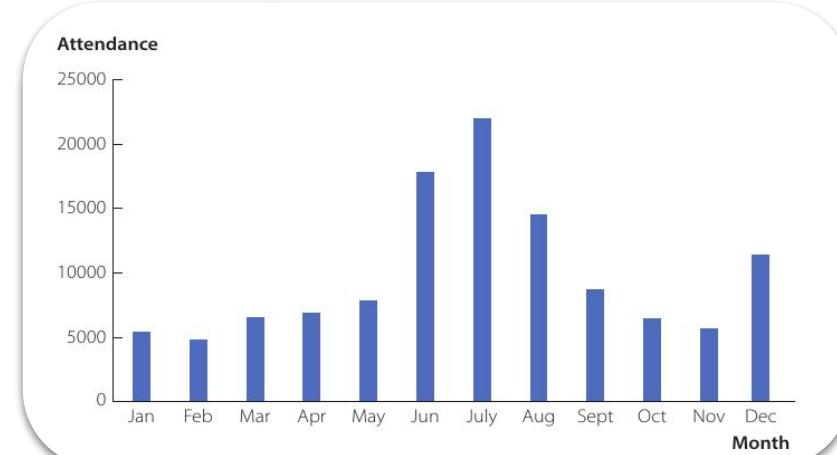
The Python logo, which is a stylized blue and yellow 'P' shape.A radar chart with five axes and colored segments (orange, yellow, green, blue, red) meeting at a central point.

Creemos nuestros gráficos

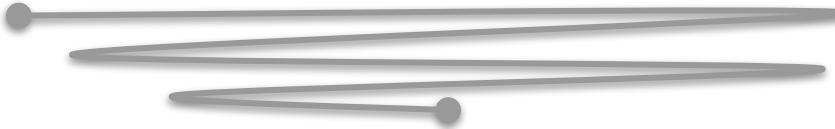
1. Importar zoo.csv

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

2. Generar el gráfico de barras (Bar Chart)



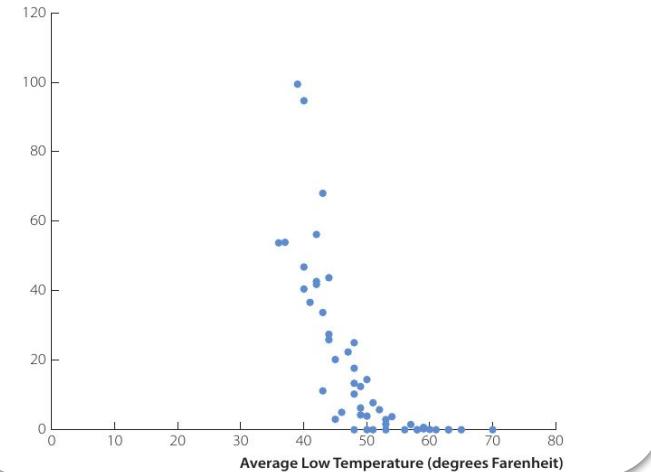
Creemos nuestros gráficos



1. Importar snow.csv

	A	B	C	D
1	City	State	Average Low Temperature	Average Snowfall
2	Atlanta	Georgia	53	2.9
3	Austin	Texas	59	0.6
4	Baltimore	Maryland	45	20.2
5	Birmingham	Alabama	53	1.6
6	Boston	Massachusetts	44	43.8
7	Buffalo	New York	40	94.7
8	Charlotte	North Carolina	49	4.3
9	Chicago	Illinois	41	36.7
10	Cincinnati	Ohio	43	11.2
11	Cleveland	Ohio	43	68.1
12	Columbus	Ohio	44	27.5
13	Dallas	Texas	57	1.5
14	Denver	Colorado	26	
15	Detroit	Michigan		
	St. Louis	Missouri		
	Portland	Connecticut		

2. Generar el gráfico de puntos (Scatter Plot)

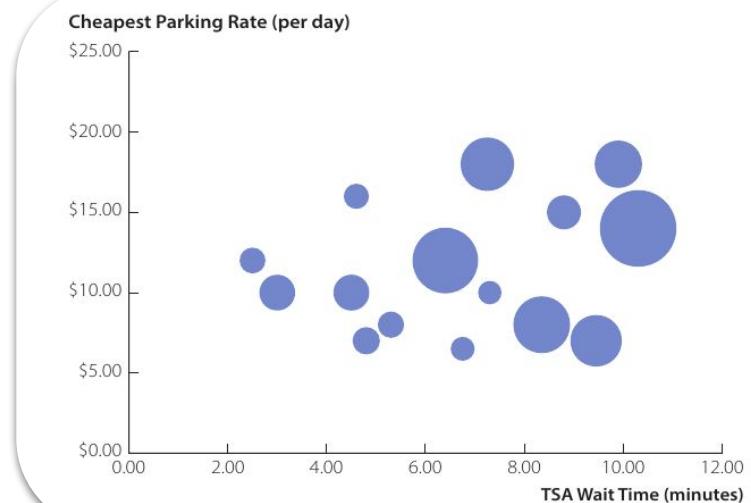


Creemos nuestros gráficos

1. Importar airport.csv

AirportCode	WaitTime	ParkingCost	AnnualEnplanements
ATL	10.3	14	49.06
CLT	9.45	7	22.19
DEN	8.35	8	27.02
AUS	5.3	8	5.8
STL	4.8	7	6.25
SMF	7.3	10	4.6
RDU	6.75	6.5	4.8
EWR	9.9	18	18.8
SFO	7.25	18	24
LAX	6.4	12	27.02
SLC	4.5	10	2.5

2. Generar el gráfico de globos/burbujas (Bubble Chart)

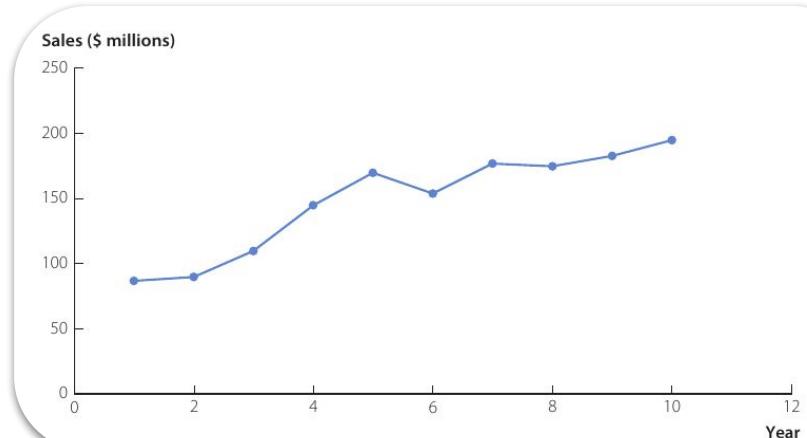


Creemos nuestros gráficos

1. Importar cheetah.csv

A	B	
1	Year	Sales (\$ millions)
2	1	87
3	2	90
4	3	110
5	4	145
6	5	170
7	6	154
8	7	177
9	8	175
10	9	183
11	10	195

2. Generar el gráfico de líneas (Line Chart)

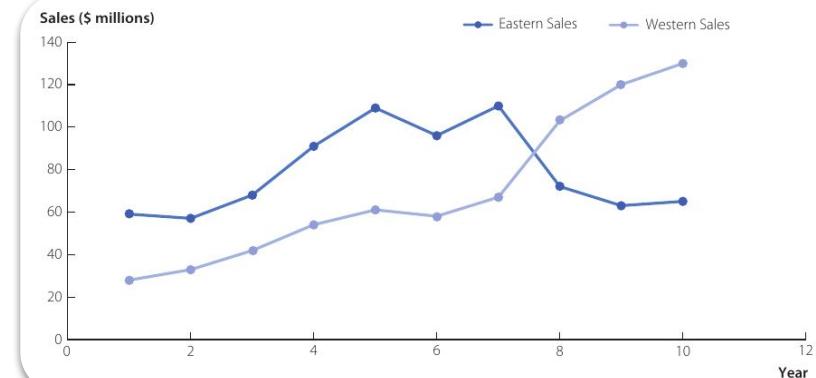


Creemos nuestros gráficos

1. Importar cheetahRegion.csv

A	B	C	D	
1	Year	Eastern Sales	Western Sales	Total Sales (\$ millions)
2	1	59	28	87
3	2	57	33	90
4	3	68	42	110
5	4	91	54	145
6	5	109	61	170
7	6	96	58	154
8	7	110	67	177
9	8	72	103	175
10	9	63		
11	10	65		

2. Generar el gráfico de líneas (Line Chart)



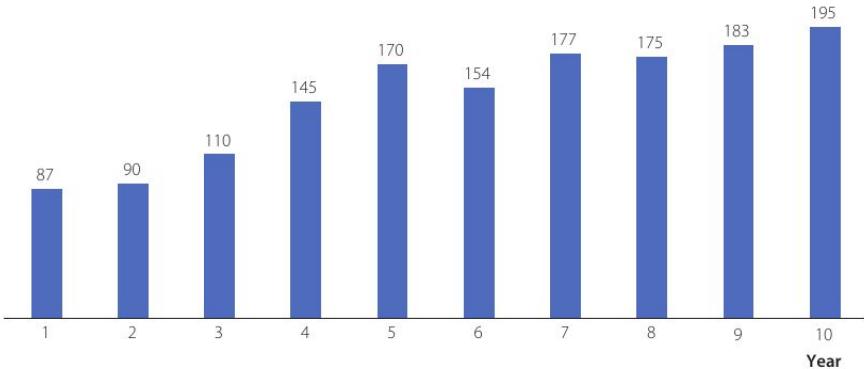
Creemos nuestros gráficos

1. Importar cheetah.csv

A	B	
1	Year	Sales (\$ millions)
2	1	87
3	2	90
4	3	110
5	4	145
6	5	170
7	6	154
8	7	177
9	8	175
10	9	183
11	10	195

2. Generar el gráfico de barras/columnas (Bar/Column Chart)

Total Sales (\$ millions)

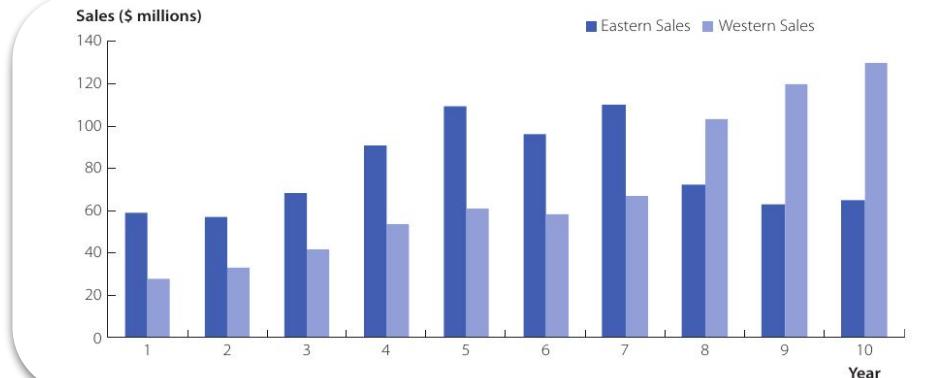


Creemos nuestros gráficos

1. Importar cheetahRegion.csv

	A	B	C	D
1	Year	Eastern Sales	Western Sales	Total Sales (\$ millions)
2	1	59	28	87
3	2	57	33	90
4	3	68	42	110
5	4	91	54	145
6	5	109	61	170
7	6	96	58	154
8	7	110	67	177
9	8	72	103	175
10	9	63		
11	10	65		

2. Generar el gráfico de barras/columnas (Bar/Column Chart)

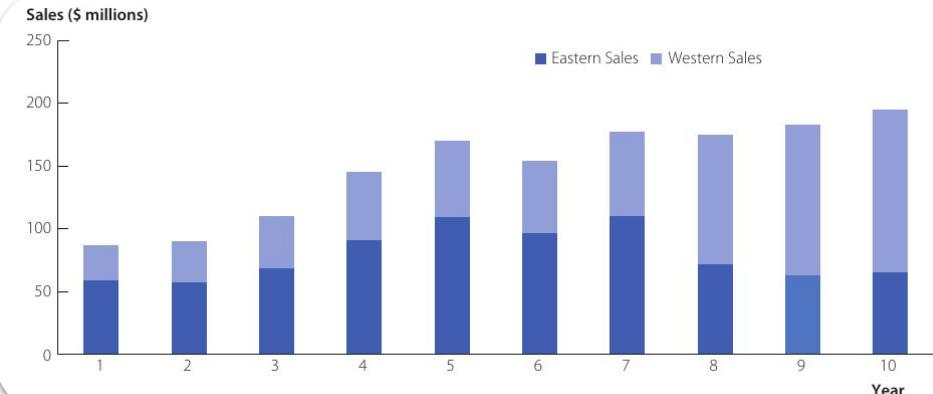


Creemos nuestros gráficos

1. Importar cheetahRegion.csv

A	B	C	D	
1	Year	Eastern Sales	Western Sales	Total Sales (\$ millions)
2	1	59	28	87
3	2	57	33	90
4	3	68	42	110
5	4	91	54	145
6	5	109	61	170
7	6	96	58	154
8	7	110	67	177
9	8	72	103	175
10	9	63		
	10	65		

2. Generar el gráfico de barras apiladas (Stacked Bar Chart)

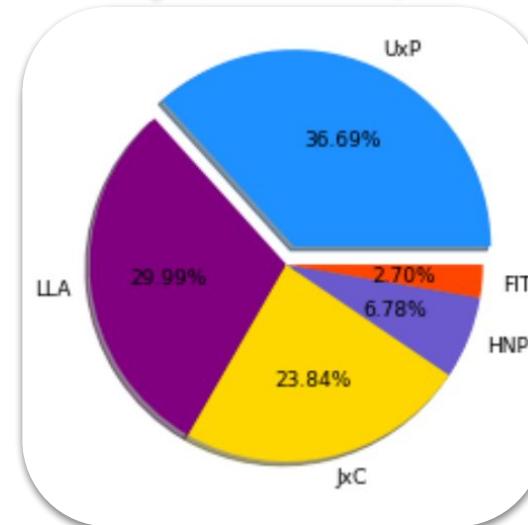


Creemos nuestros gráficos

1. Importar votacionGeneral.csv

Agrupacion	Alias	Porcentaje
Unión por la Patria	UxP	36,69
La libertad Avanza	LLA	29,99
Juntos por el Cambio	JxC	23,84
Hacemos por Nuestro País	HNP	6,78
Frente de Izquierda y de Trabajadores - Unidos	FIT	2,7

Generar el gráfico de torta (Pie Chart)



Ejercicios



Consigna.

1. Sean los siguientes datos correspondientes a los porcentajes de ventas de una compañía según la región
2. Generar un gráfico para representarlos gráficamente
3. Analizar los resultados obtenidos
4. 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?

Region	Porcentaje
Este	28
Norte	14
Sur	36
Oeste	22

Ejercicios



Consigna.

1. Sean los siguientes datos correspondientes a los precios del biodiesel en distintos períodos en la Argentina
(se encuentran subidos en el campus)
2. Generar un gráfico para representarlos gráficamente
3. Analizar los resultados obtenidos
4. 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?

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

Ejercicios



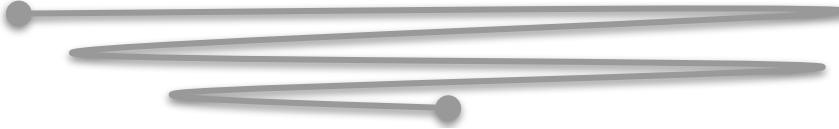
Consigna.

1. Sean los siguientes datos correspondientes a poseedores de teléfonos
(se encuentran subidos en el campus)

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

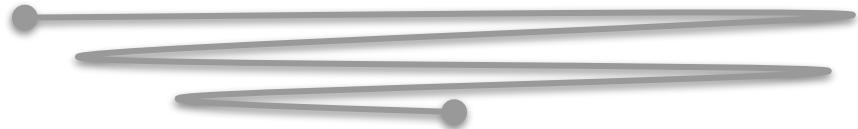
2. Generar un gráfico para representarlos gráficamente
3. Analizar los resultados obtenidos
4. Discutir con el resto de la clase
 - a. ¿Cuál fue su objetivo: Explorar, Explicar, Otro?
 - b. ¿Qué tipos de variables estaban en juego?
 - c. ¿Qué tipo de gráfico decidió utilizar?
 - d. ¿Qué resultados obtuvo?
 - 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."

Visualización



Distribución de los Datos

Visualización - Distribución de los Datos



- No es intuitivo
- Aumenta carga cognitiva de la audiencia

Edades de los atletas olímpicos de USA en los últimos 4 torneos de verano
(Bordes: izquierdo -> mínimo; derecho -> máximo)

FIGURE 5.1 Overlapping Range Bar Chart For Ages Of U.S. Olympic Gymnasts

■ Male ■ Female ■ Both

Age Range of U.S. Gymnasts in the Four Most Recent Summer Games



1. ¿En qué rango de edades 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?

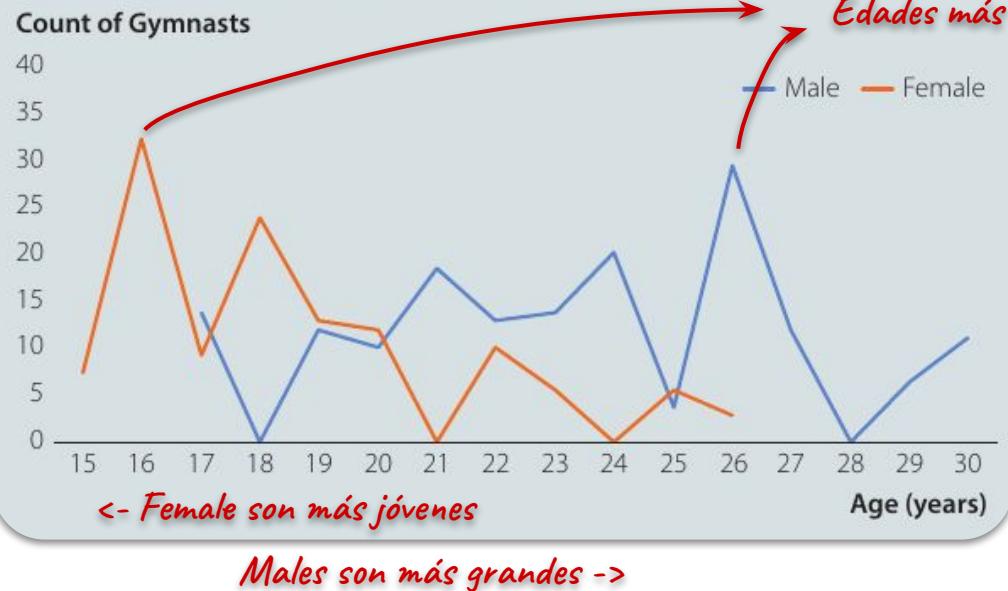
¡No hay información sobre cómo se distribuyen lxs gimnastas masculinxs y femeninxs en sus respectivos rangos!

Visualización - Distribución de los Datos

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

FIGURE 5.2 Frequency Polygon for U.S. Olympic Gymnasts

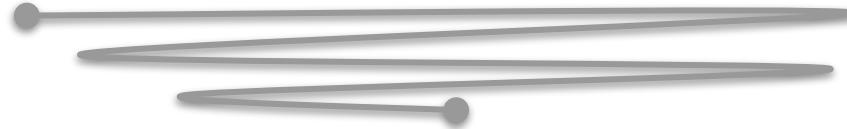
Age Distribution of U.S. Olympic Gymnasts



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 gimnastas 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 ...

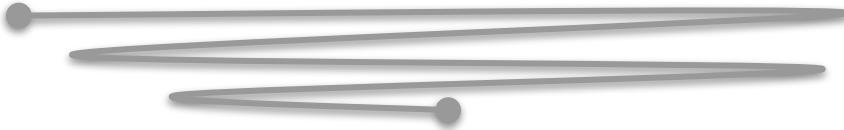
Describe qué valores se observaron y con qué frecuencia esos valores aparecen en dichos datos

Distribución
de
frecuencia



- a. Variable categórica
(etiquetas que no pueden manipularse aritméticamente)
- b. Variable cuantitativa
(valores numéricos que pueden manipularse aritméticamente)

Visualización - Distribución de los Datos



a. Variables Categóricas

Visualización - Distribución de los Datos - Categóricos

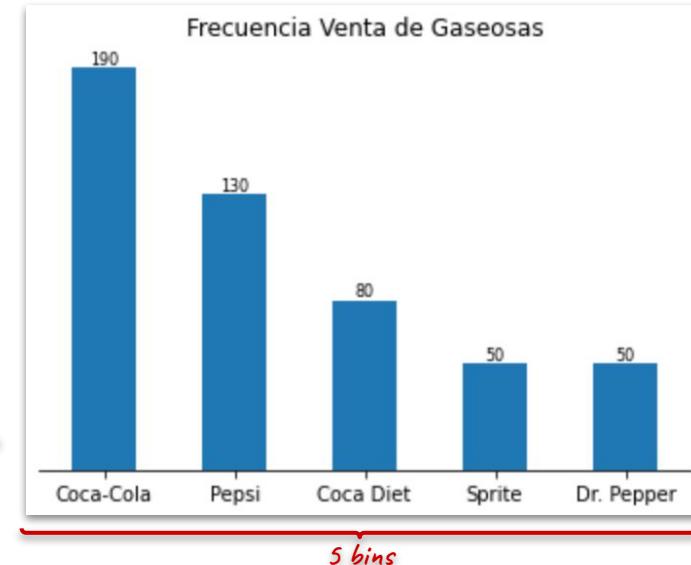
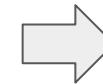
Una distribución de frecuencia (de variable categóricas) es un resumen de datos que muestra el número (frecuencia) de observaciones en cada una de las clases (no superpuestas), denominadas **bins**.

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

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



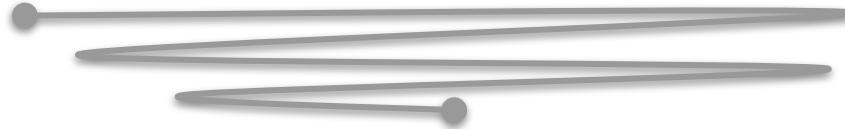
Creemos nuestro gráfico



You can use text.
plt.text(x, y,
z)
t
g- python™ give
axis. the pos. of
your text will be



Visualización - Distribución de los 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 es la Frecuencia relativa multiplicada por 100

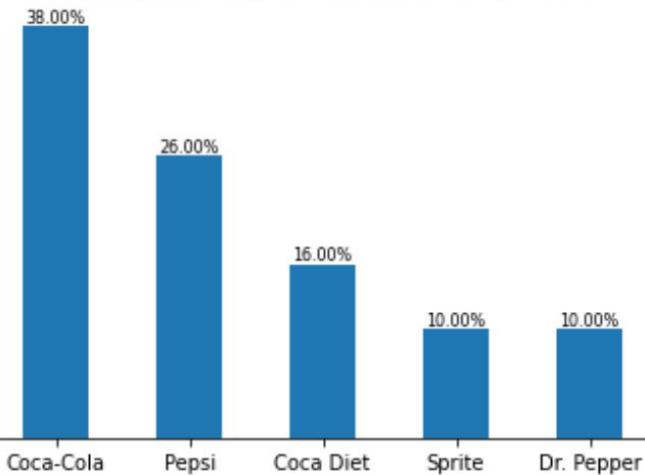
Visualización - Distribución de los Datos - Categóricos

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

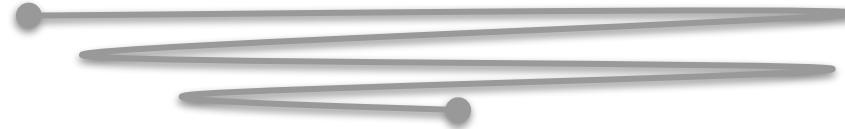
Frecuencia relativa

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

Frecuencia Porcentual de Venta de Gaseosas



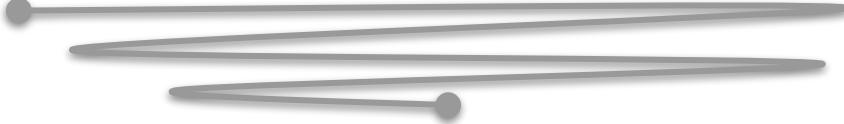
Creemos nuestro gráfico



You can use text.
plt.text(x, y, text)
z
t
g- python™ give
axis. the position of
your text will be

A circular icon containing the Python logo, which is a stylized blue and yellow 'P'. Below the logo, the word "python" is written in a lowercase, sans-serif font, followed by a trademark symbol.A circular icon showing a polar plot with radial axes. Several colored arrows (orange, yellow, green, blue) are radiating from the center, pointing towards the outer rings of the grid.

Visualización - Distribución de los 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

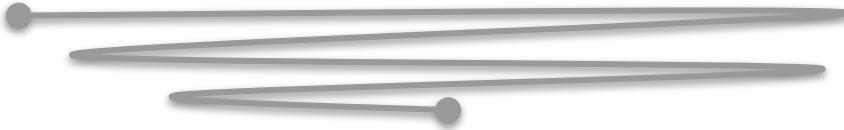
¿Cómo dividirían este total entre los distintos tipos de gaseosas individuales?

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.

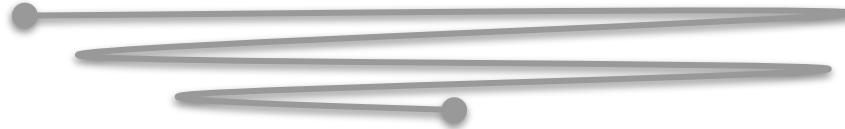
Compras_gaseosas	
Coca-Cola	0.38
Pepsi	0.26
Coca Diet	0.16
Dr. Pepper	0.10
Sprite	0.10

Visualización - Distribución de los Datos



b. Variables Continuas

Visualización - Distribución de los Datos - Continuos



¿Cuál es la dificultad para obtener la distribución de frecuencia en variables continuas?

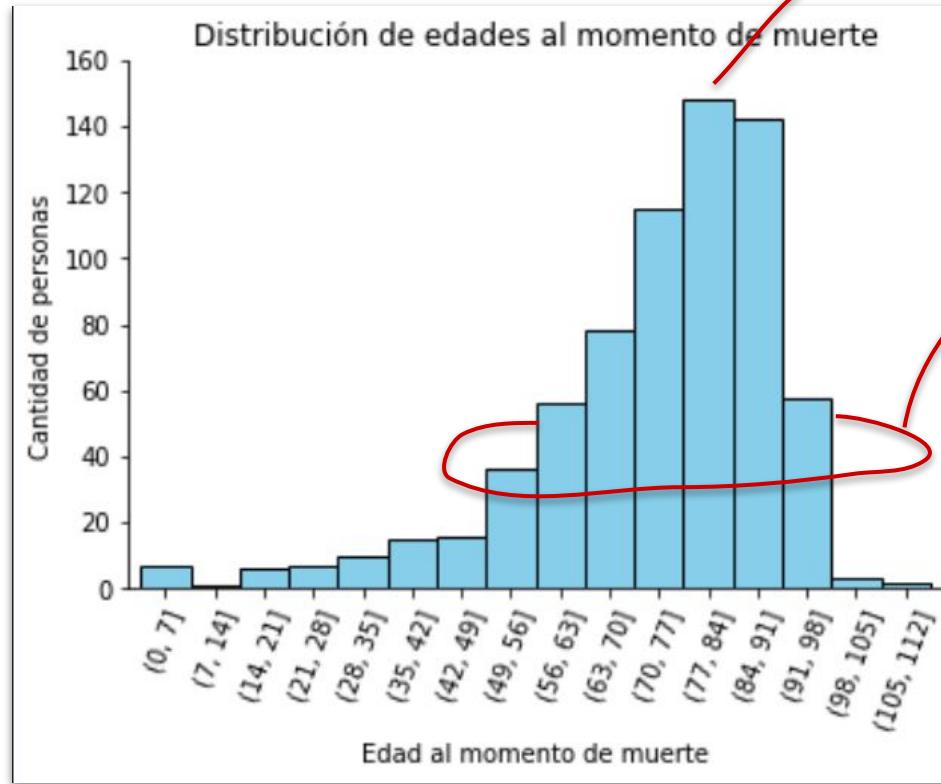
Pensar Sexo (categórica, asumir 2 valores posibles) vs. Peso (continua, asumir 3 decimales)

Solución:

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

Visualización - Distribución de los Datos - Continuos

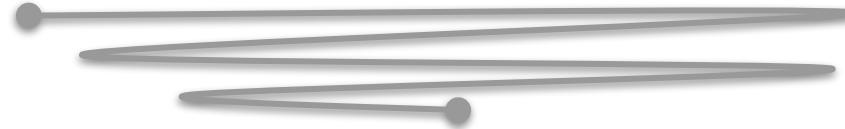
AgeAtDeath
81
64
88
85
96
101
87
8



Edad más frecuente de muerte

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

Creemos nuestro gráfico



You can use text.
plt.text(x, y,
z)
t
g- python™ give
axis. the pos. of
your text will be



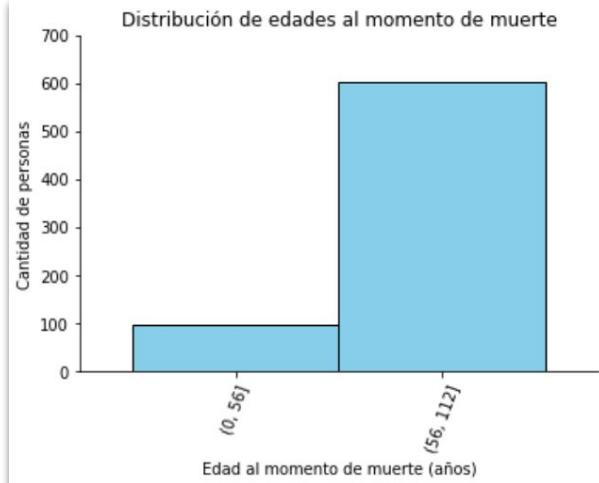
Visualización - Distribución de los Datos - Continuos

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

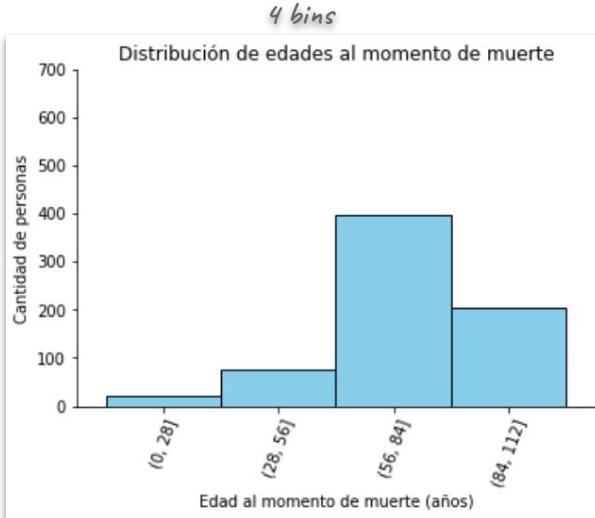
Tarea: Definir ...

1. La cantidad de bins
2. El ancho (rango numérico) de cada bin
3. El rango total que abarca el conjunto de bins

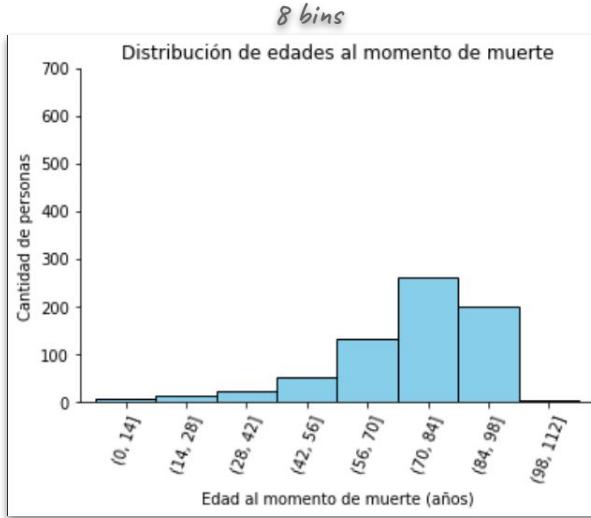
2 bins



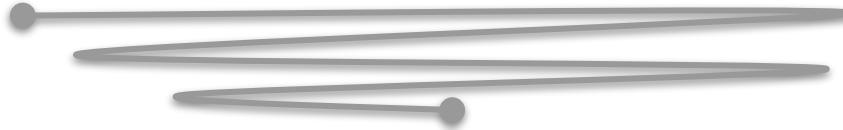
4 bins



8 bins



Visualización - Distribución de los 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.

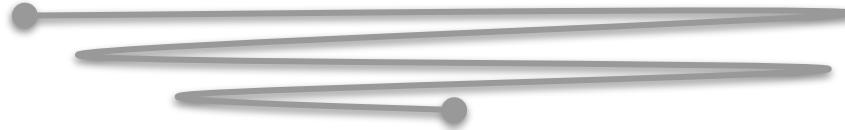
La elección de la cantidad de bins es subjetiva, depende del tema y el objetivo del análisis.

Recomendación: utilizar de 5 a 20 bins ...

Pocas observaciones -> 5 o 6 bins

Muchas observaciones -> Más bins.

Visualización - Distribución de los Datos - Continuos



2. Ancho de bins ...

- Tomar anchos distintos para cada bin puede llevar a decisiones equivocadas*

Recomendación: utilizar bins del mismo ancho

Visualización - Distribución de los Datos - Continuos



3. Rango de valores de bins ...

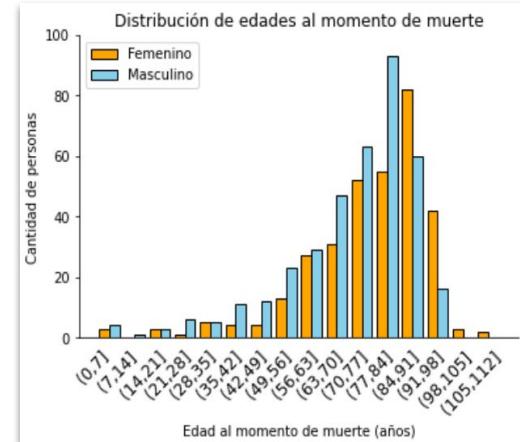
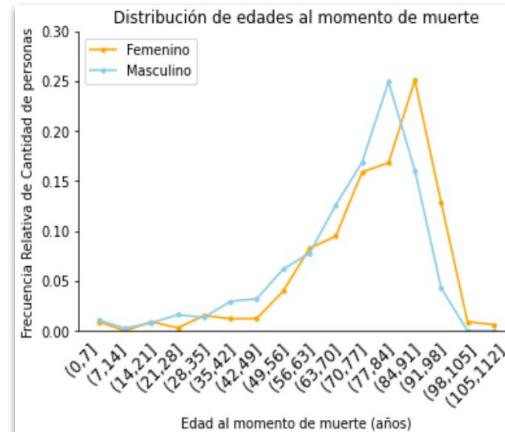
- *Todas las observaciones deberian caer dentro de un bin*
- *Los bins no deben superponerse*
- *Tener cuidado con los extremos*

Recomendación: utilizar rangos de bins que cumplan con lo anterior

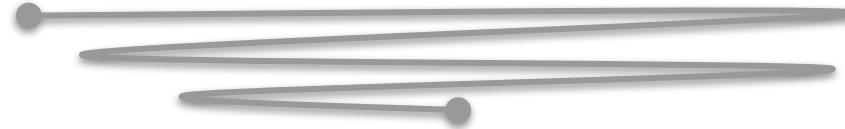
Visualización - Distribución de los Datos - Continuos

¿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	



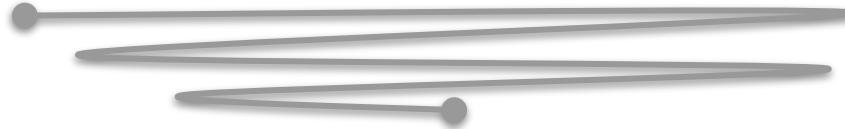
Creemos nuestro gráfico



You can use text.
plt.text(x, y,
z)
t
g- python™ give
axis. the pos. of
your text will be

A circular icon containing the Python logo, which is a stylized blue and yellow 'P'. Below the logo, the word "python™" is written in a lowercase sans-serif font.A circular icon showing a polar plot with radial axes and several colored vectors (orange, yellow, green, blue) originating from the center, illustrating a complex number or vector magnitude and angle.

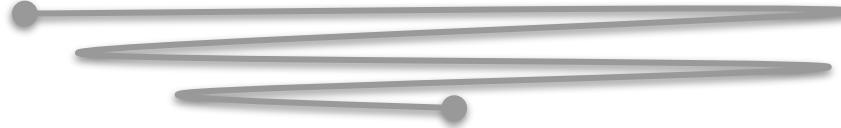
Ejercicios



Consigna.

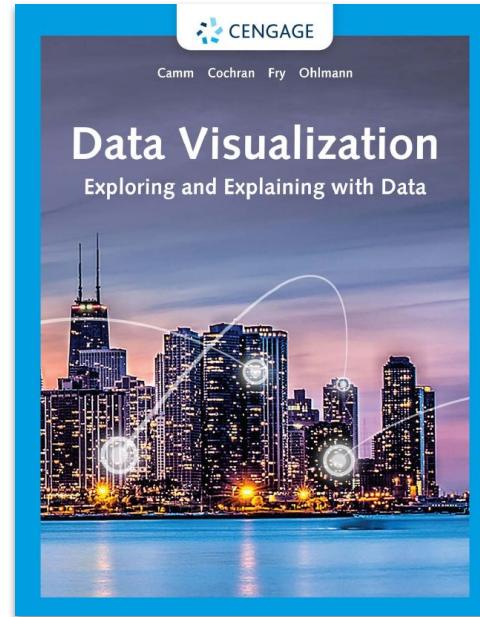
1. Sean los datos correspondientes a las propinas de un bar
2. Generar un gráfico para analizar la distribución de la propina en función del:
 - a. sexo
 - b. dia de la semana
3. Comentar los resultados obtenidos

Cierre



1. *Exploración y Explicación*
2. *Distintas maneras de visualizar y explorar datos*

Bibliografia



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