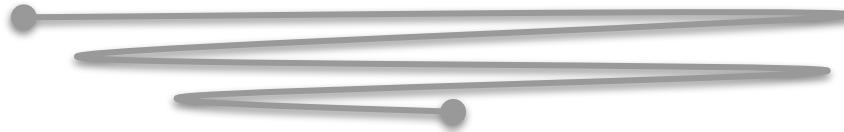


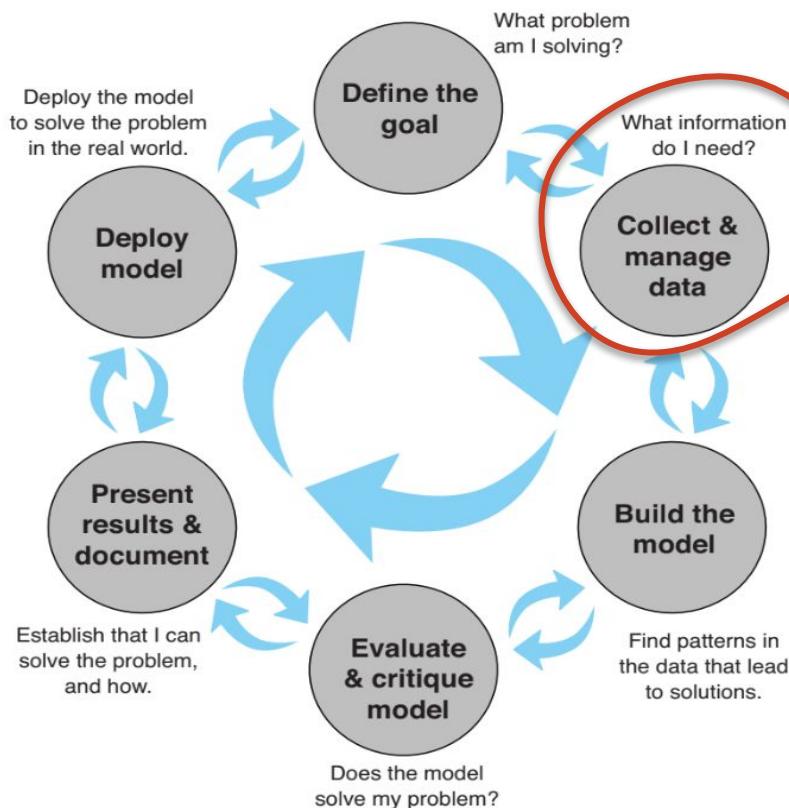
Laboratorio de Datos



Visualización y Análisis Exploratorio de Datos - Parte 02



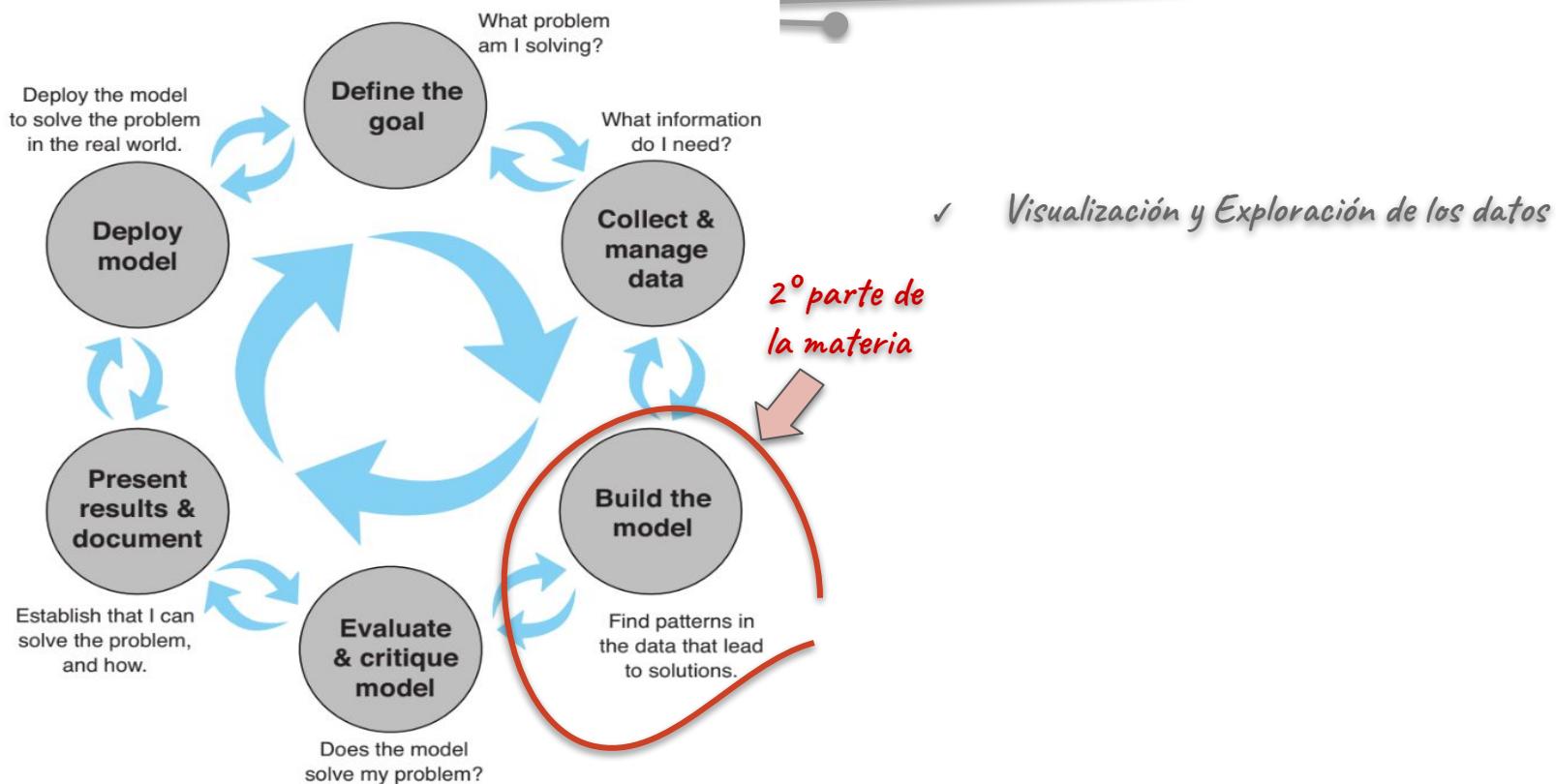
Recorrido de la materia (hasta ahora)



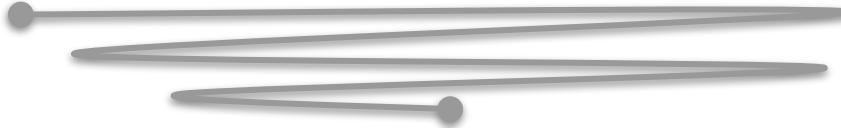
1º parte de
la materia

- ✓ 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)

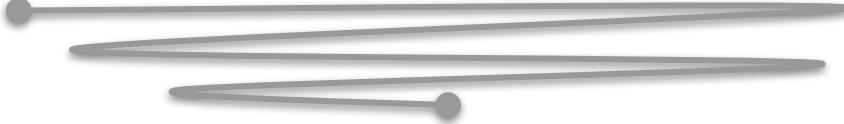


Repaso clase anterior



1. *Exploración y Explicación*
2. *Distintas maneras de visualizar y explorar datos*
3. *Ejemplos*
4. *Distribución de Datos (Histogramas de variables categóricas y continuas)*

Visualización - Análisis Estadístico

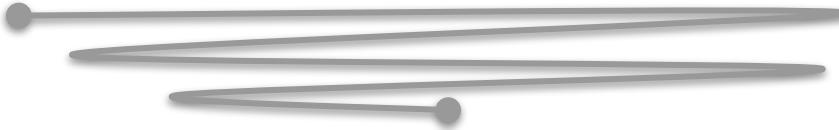


Pensar un único valor que mejor caracterice el precio de venta de estas propiedades ... (2 min.)

PrecioDeVenta
\$108.000
\$138.000
\$138.000
\$142.000
\$186.000
\$199.500
\$208.000
\$254.000
\$254.000
\$257.500
\$298.000
\$456.400

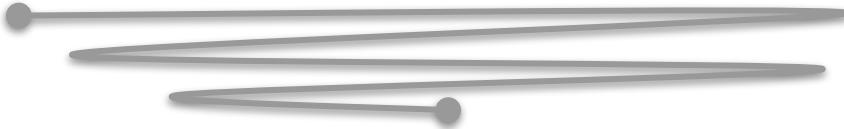
¿Cómo lo obtuvieron?

Visualización - Análisis Estadístico



Medidas de Tendencia

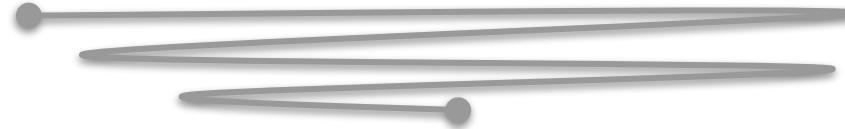
Visualización - Análisis Estadístico - Medidas de Tendencia



Una medida de tendencia (central) es un valor único asociado a una variable para caracterizar de alguna manera el conjunto completo de valores

- Existen distintas medidas
- Cada una posee ventajas y desventajas relativas respecto a las otras

Visualización - Análisis Estadístico - Medidas de Tendencia



Media (o valor promedio) es la sumatoria de todos los datos dividida la cantidad total de datos

PrecioDeVenta
\$108.000
\$138.000
\$138.000
\$142.000
\$186.000
\$199.500
\$208.000
\$254.000
\$254.000
\$257.500
\$298.000
\$456.400

$$\text{Media} = \frac{\$108.000 + \$138.000 + \dots + \$456.400}{12} = \$219.950$$

Visualización - Análisis Estadístico - Medidas de Tendencia



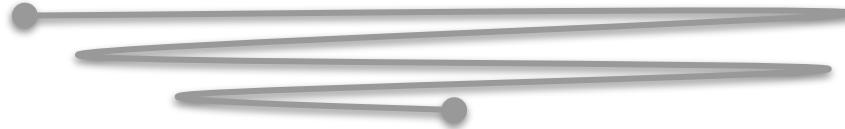
Mediana es el número del medio; se encuentra al ordenar todos los valores y elegir el que está en el medio
(o si hay dos números en el medio, tomar el promedio de esos dos números)

PrecioDeVenta
\$108.000
\$138.000
\$138.000
\$142.000
\$186.000
\$199.500
\$208.000
\$254.000
\$254.000
\$257.500
\$298.000
\$456.400

Ya
están
ordenados

$$\text{Mediana} = \frac{\$199.500 + \$208.000}{2} = \$203.750$$

Visualización - Análisis Estadístico - Medidas de Tendencia



Media (promedio) -> Sí, influenciada por valores atípicos (valores extremadamente chicos/grandes)

Mediana

-> No influenciada por valores atípicos (su cálculo es robusto)

PrecioDeVenta
\$108.000
\$138.000
\$138.000
\$142.000
\$186.000
\$199.500
\$208.000
\$254.000
\$254.000
\$257.500
\$298.000
\$456.400

$$\underbrace{\$219.950}_{\text{Media}} > \underbrace{\$203.750}_{\text{Mediana}}$$

Aumenta la media
pero no la mediana

Valor
extremo

Si reemplazáramos los \$456.400 por \$1,5 millones
(media = \$306.916,67; mediana = \$203.750)
la mediana permanecería sin cambios



La mediana da una mejor
idea del precio de venta

Siempre que un conjunto de datos contiene valores extremos, la
mediana es la medida preferida de tendencia central
(en particular para conjuntos de datos con pocas observaciones)

Visualización - Análisis Estadístico - Medidas de Tendencia

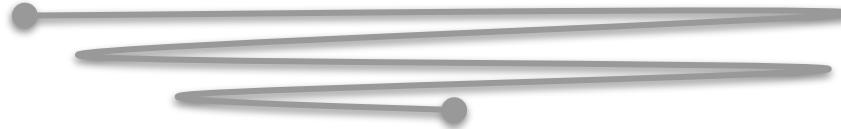
Moda es el número más frecuente, es decir, el número que se repite el mayor número de veces (en caso de empate, puede existir más de una moda; en caso de no existir repeticiones los datos no tienen moda)

PrecioDeVenta	
se repite 2 veces	\$108.000
	\$138.000
	\$138.000
	\$142.000
	\$186.000
	\$199.500
	\$208.000
	\$254.000
se repite 2 veces	\$254.000
	\$257.500
	\$298.000
	\$456.400

$$\text{Moda} = \{ \$138.000; \$254.000 \}$$

- Útil para variables que tienen pocos valores distintos
- Variables con muchos valores distintos (Ej. tiempos de maratonistas en una carrera)
 - Es posible que la moda no exista, ¿por qué?
 - Alternativa: construir histograma y aplicar la noción de moda para referirse al bin con mayor cantidad de observaciones.

Visualización - Análisis Estadístico - Medidas de Dispersion



Consigna. Sean ...

Notas Estudiante A: 4; 5; 7; 7; 7; 9; 10

Notas Estudiante B: 7; 7; 7; 7; 7; 7; 7

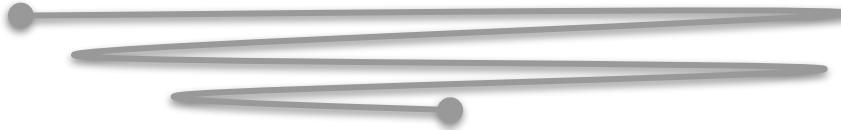
Calcular la Media, Mediana y Moda para cada uno de los Estudiantes

Respuestas.

	Estudiante A	Estudiante B
Media	7	7
Mediana	7	7
Moda	7	7

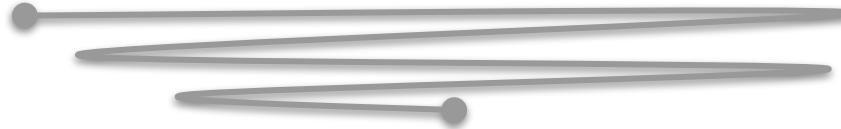
- *Medidas de tendencia central no describen de qué manera varian los valores*
- *Es necesario un valor permita caracterizar a la “dispersión” de los valores ¿Cuál?*

Visualización - Análisis Estadístico



Medidas de Dispersion

Visualización - Análisis Estadístico - Medidas de Dispersion



Rango es la diferencia numérica entre el valor máximo y el valor mínimo

PrecioDeVenta
\$108.000
\$138.000
\$138.000
\$142.000
\$186.000
\$199.500
\$208.000
\$254.000
\$254.000
\$257.500
\$298.000
\$456.400

Mínimo

$$\text{Rango} = \$456.400 - \$108.000 = \$348.400$$

Máximo

¿Problemas?

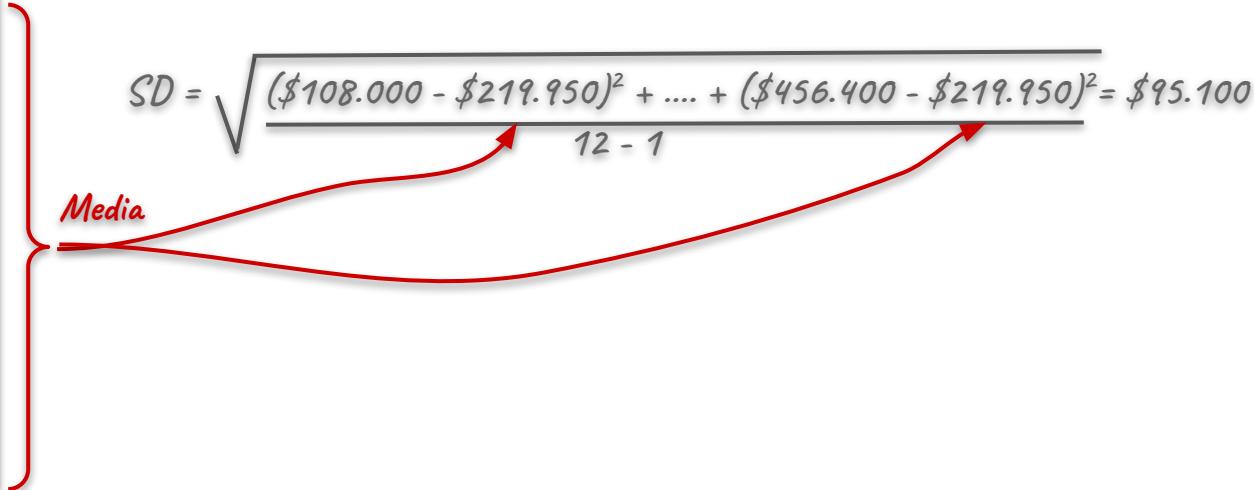
- Se basa sólo en 2 valores (máximo y mínimo)
- Influenciable por valores extremos

Visualización - Análisis Estadístico - Medidas de Dispersion



Desviación Estándar representa cuánto se apartan los valores del valor promedio

PrecioDeVenta
\$108.000
\$138.000
\$138.000
\$142.000
\$186.000
\$199.500
\$208.000
\$254.000
\$254.000
\$257.500
\$298.000
\$456.400



Visualización - Análisis Estadístico - Medidas de Dispersion



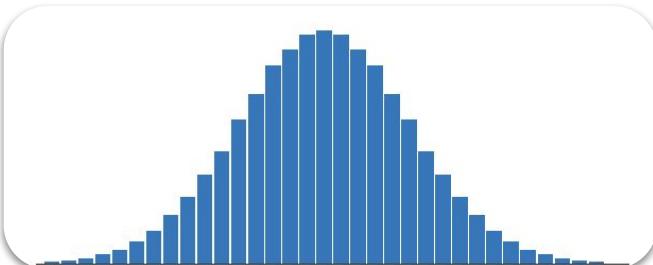
Desviación Estándar representa cuánto se apartan los valores del valor promedio

- Medida confiable cuando el histograma tiene forma de campana simétrica
- En estos casos la variabilidad nos permite describir los datos usando intervalos ...

- el 68% de los valores están en [media - 1 SD; media + 1 SD]
- el 95% de los valores están en [media - 2 SD; media + 2 SD]
- > 99% de los valores están en [media - 3 SD; media + 3 SD]

¿Problemas?

- No es confiable para distribuciones asimétricas
- Influenciable por valores extremos



Visualización - Análisis Estadístico - Medidas de Dispersion



Percentil es el valor que divide a una lista ordenada de datos de forma que un porcentaje de los datos sea inferior a dicho valor

¿Cómo se calcula el valor del p -ésimo percentil?

1. Calcular su posición entre el conjunto de valores ordenados

Ejemplo. ¿percentil 25? $\rightarrow \frac{25 \times (12 + 1)}{100} = 3,25$

entre posiciones 3 y 4

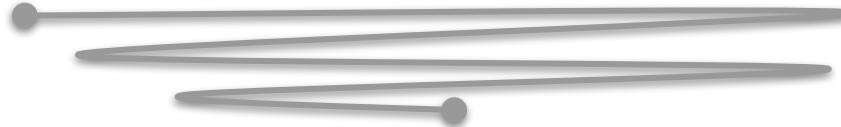
2. Realizar la interpolación necesaria

$$\$138.000 + (3,25 - 3) \times (\$142.000 - \$138.000) = \$139.000$$

PrecioDeVenta
\$108.000
\$138.000
\$138.000
\$142.000
\$186.000
\$199.500
\$208.000
\$254.000
\$254.000
\$257.500
\$298.000
\$456.400

} 25% de los datos

Visualización - Análisis Estadístico - Medidas de Dispersion



Percentil es el valor que divide a una lista ordenada de datos de forma que un porcentaje de los datos sea inferior a dicho valor

¿Cómo calcular el valor del p -ésimo percentil?

1. Calcular su posición entre el conjunto de valores ordenados

Ejemplo. ¿percentil 50? $\rightarrow \frac{50 \times (12 + 1)}{100} = 6,5$

entre posiciones 6 y 7

2. Realizar la interpolación necesaria

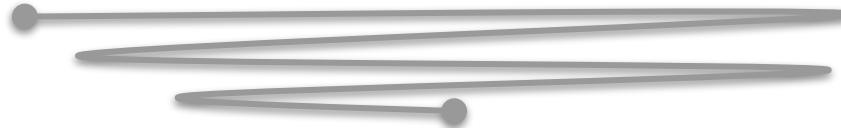
$$\$199.500 + (6,5 - 6) \times (\$208.000 - \$199.500) = \$203.750$$

¡Coincide con la mediana!

PrecioDeVenta
\$108.000
\$138.000
\$138.000
\$142.000
\$186.000
\$199.500
\$208.000
\$254.000
\$254.000
\$257.500
\$298.000
\$456.400

50% de los datos

Visualización - Análisis Estadístico - Medidas de Dispersion



Percentil es el valor que divide a una lista ordenada de datos de forma que un porcentaje de los datos sea inferior a dicho valor

¿Cómo calcular el valor del p -ésimo percentil?

1. Calcular su posición entre el conjunto de valores ordenados

Ejemplo. ¿percentil 75? -> $\frac{75 \times (12 + 1)}{100} = 9,75$ ————— entre posiciones 9 y 10

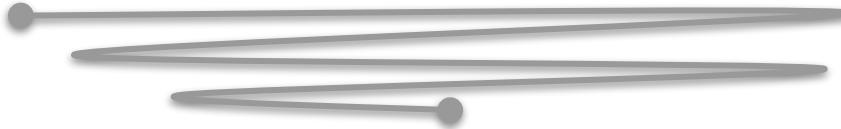
2. Realizar la interpolación necesaria

$$\$254.000 + (9,75 - 9) \times (\$257.500 - \$254.000) = \$256.625$$

PrecioDeVenta
\$108.000
\$138.000
\$138.000
\$142.000
\$186.000
\$199.500
\$208.000
\$254.000
\$254.000
\$257.500
\$298.000
\$456.400

75% de los datos

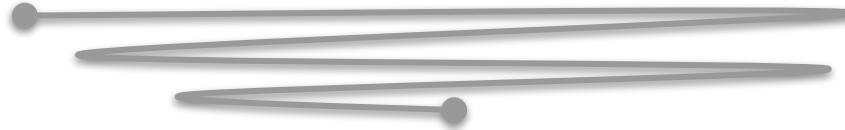
Visualización - Análisis Estadístico - Medidas de Dispersion



Percentil es el valor que divide a una lista ordenada de datos de forma que un porcentaje de los datos sea inferior a dicho valor

- Se puede calcular un percentil para cualquier valor entre 0% y 100%
- Percentiles más comunes: 25, 50 y 75 (primer cuartil, segundo cuartil y tercer cuartil)
- Rango intercuartil (IQR). La diferencia entre el tercer y el primer cuartil (los percentiles 75 y 25)
- IQR abarca el 50% medio de la distribución de los valores y se utiliza como medida de variación
- Ventajas de percentiles y el rango intercuartil sobre el rango y la desviación estándar
 - percentiles no requieren que la distribución de una variable tenga forma de campana
 - valores extremos no distorsionan el valor de los percentiles

Visualización - Análisis Estadístico - Medidas de Dispersion



Consigna. Sean ...

Notas Estudiante A: 4; 5; 7; 7; 7; 9; 10

Notas Estudiante B: 7; 7; 7; 7; 7; 7; 7

Calcular la Media, Mediana y Moda para cada uno de los Estudiantes

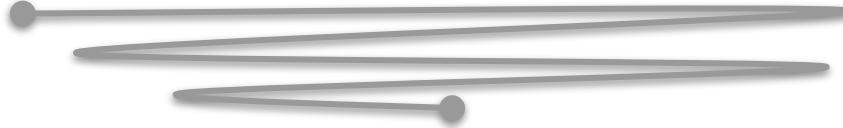
Calcular el rango, desvío standar, cuartiles e IQR

Respuestas.

	Estudiante A	Estudiante B
Media	7	7
Mediana	7	7
Moda	7	7

	Estudiante A	Estudiante B
Rango	6	0
STD	2,08	0,00
1Q	5	7
2Q	7	7
3Q	9	7
IQR	2	0

Visualización - Distribución

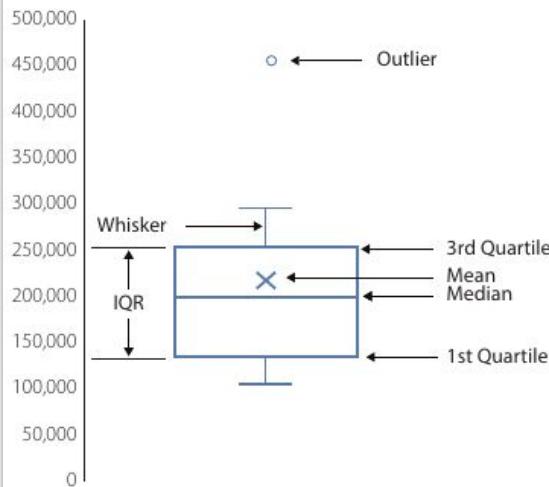


Análisis Estadístico - Boxplot

Visualización - Análisis Estadístico - Boxplot

Boxplot es un resumen gráfico de la distribución de los datos. Se basa en los cuartiles.

Home Selling Prices (\$)



Caja (Box)

$$3\text{er. cuartil} = \text{Percentil } 75 = \$256.625$$

$$2\text{do. cuartil} = \text{Percentil } 50 = \$203.750 \text{ (Mediana.)}$$

$$1\text{er. cuartil} = \text{Percentil } 25 = \$139.000$$

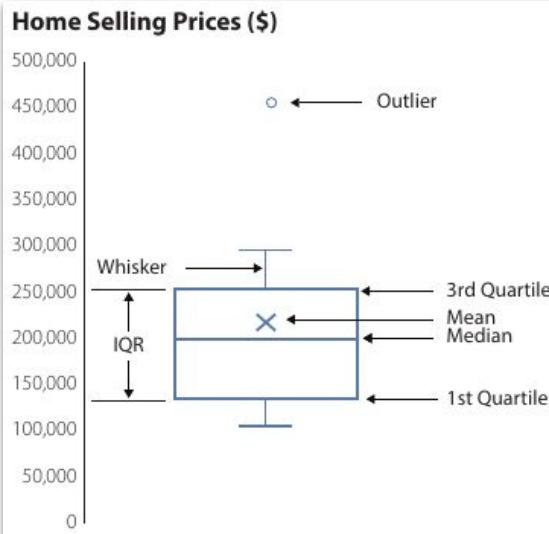
$$IQR = 3\text{er.} - 1\text{er cuartil} = \$117.625$$

$$\text{Media} = \$219.950$$

PrecioDeVenta
\$456.400
\$298.000
\$257.500
\$254.000
\$254.000
\$208.000
\$199.500
\$186.000
\$142.000
\$138.000
\$138.000
\$108.000

Visualización - Análisis Estadístico - Boxplot

Boxplot es un resumen gráfico de la distribución de los datos. Se basa en los cuartiles.



Bigotes (Whisker)

$$\text{Límite Sup.} = \text{3er. cuartil} + 1,5 \times \text{IQR} = \$433,062.5$$

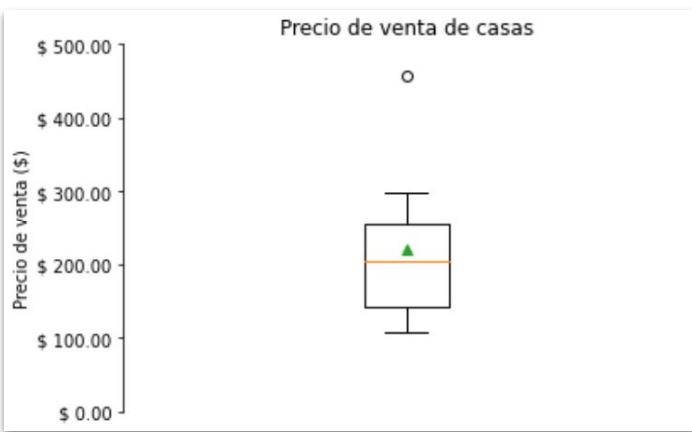
$$\text{Límite Inf.} = \text{1er. cuartil} - 1,5 \times \text{IQR} = \$-37.437$$

PrecioDeVenta
\$456.400
\$298.000
\$257.500
\$254.000
\$254.000
\$208.000
\$199.500
\$186.000
\$142.000
\$138.000
\$138.000
\$108.000

← Outlier
← Bigote Sup.
← Bigote inf.

Visualización - Análisis Estadístico - Boxplot

Boxplot es un resumen gráfico de la distribución de los datos. Se basa en los cuartiles.



Bigotes (Whisker)

$$\text{Límite Sup.} = \text{3er. cuartil} + 1,5 \times \text{IQR} = \$433,062.5$$

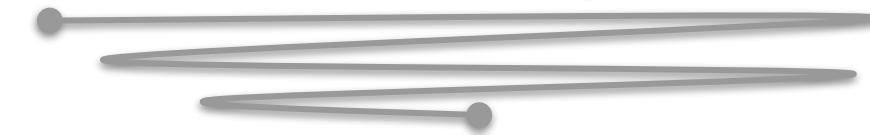
$$\text{Límite Inf.} = \text{1er. cuartil} - 1,5 \times \text{IQR} = \$-37.437$$

PrecioDeVenta
\$456.400
\$298.000
\$257.500
\$254.000
\$254.000
\$208.000
\$199.500
\$186.000
\$142.000
\$138.000
\$138.000
\$108.000

← Outlier
← Bigote Sup.

← Bigote inf.

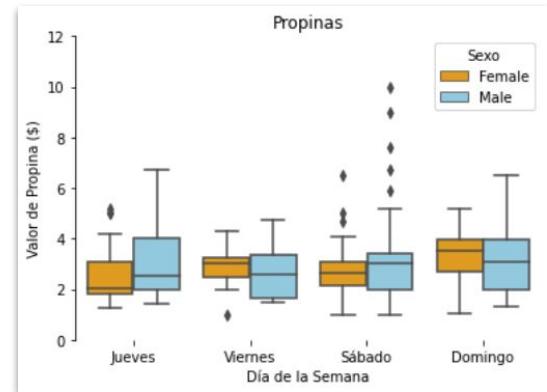
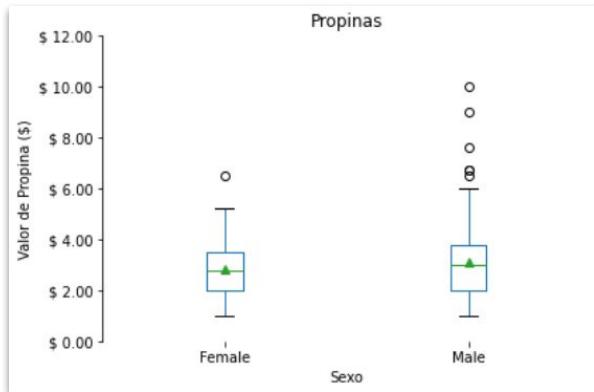
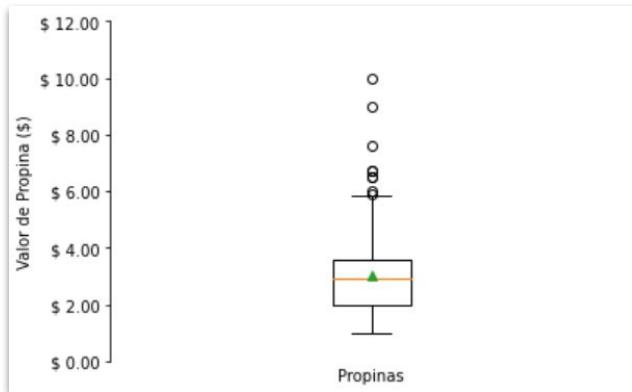
Creemos nuestros gráficos



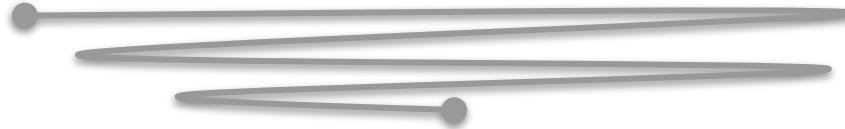
You can use text.
plt.text(x, y, text)
z
t
g- python™ give
axis. the position of
your text will be

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 featuring a radar chart or sunburst diagram. It has a central point with several radial lines extending outwards. Each segment is filled with a different color: orange, yellow, green, blue, and red. The background of the chart is light grey with concentric circles.

Visualización - Análisis Estadístico - Boxplot



Ejercicios



Consigna.

1. Sean los siguientes datos correspondientes a las propinas
2. Generar un gráfico de boxplot del porcentaje de propina recibido ($\text{tip}/\text{total_bill}$)
3. Analizar los resultados obtenidos
4. Estudiar si hay diferencias en función del sexo y el sexo+día
5. Discutir con el resto de la clase los resultados obtenidos

Visualización - Distribución

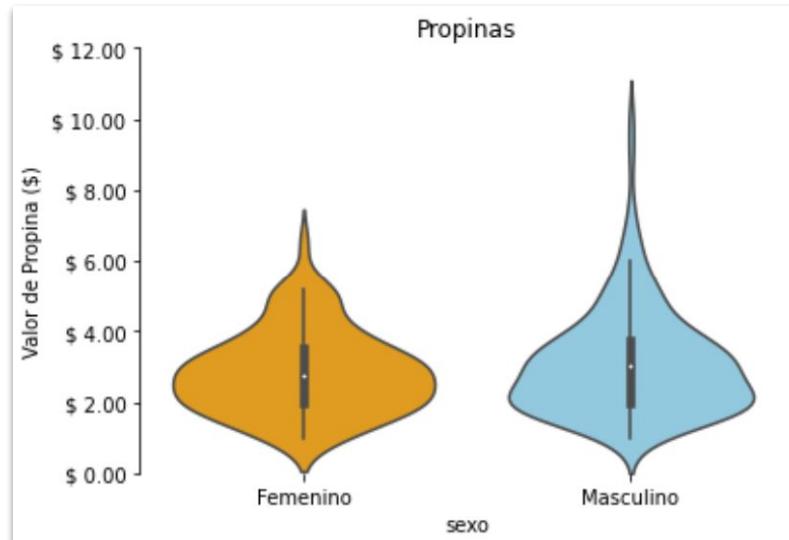


Análisis Estadístico - Violinplot

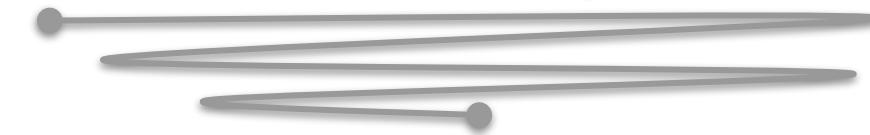
Visualización - Análisis Estadístico - Violinplot



Violinplot es similar al boxplot, salvo que muestra también la densidad de probabilidad de los datos, generalmente suavizada por un estimador de densidad kernel.



Creemos nuestros gráficos



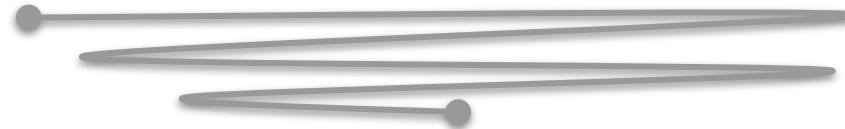
You can use text.
plt.text(x, y, text)
z
t
g- python™ give
axis. the position of
your text will be

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.

Tips para mejorar la visualización de datos



Tips para mejorar la visualización de datos - La memoria



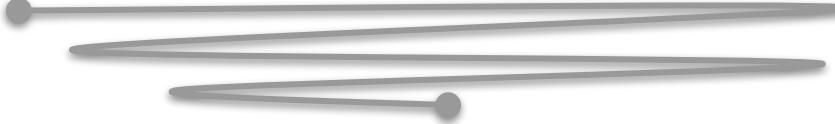
The graphic features two books. On the left, a red book is titled "Memory Systems". On the right, a yellow book is titled "Bad Recipes". Below the books, the word "Psychology" is written in large, bold, black letters. At the bottom left, there is a small icon of an open book with a white star in the center.

Memory Systems

Psychology

Bad Recipes

Tips para mejorar la visualización de datos - La memoria



Memoria sensorial	
<i>Definición</i>	Procesos cerebrales que interpretan estímulos por períodos mucho más breves que la memoria de corto plazo
<i>Sistemas derivados</i>	<ul style="list-style-type: none">- Memoria icónica- Memoria ecoica- Memoria olfativa- Memoria gustativa- Memoria háptica
<i>Tiempo de permanencia de los datos</i>	Breve (milésimas de seg)
<i>Ejemplos</i>	Percibir un sonido en medio de una multitud

Fuente:<https://www.diferenciador.com/tipos-de-memoria/>

Tips para mejorar la visualización de datos - La memoria

	Memoria sensorial	Memoria a corto plazo
Definición	Procesos cerebrales que interpretan estímulos por períodos mucho más breves que la memoria de corto plazo	Procesos cerebrales encargados de interpretar los estímulos y conservar esa información durante un tiempo breve
Sistemas derivados	- Memoria icónica - Memoria ecoica - Memoria olfativa - Memoria gustativa - Memoria háptica	- Sistema ejecutivo - Almacén episódico - Bucle fonológico - Agenda visoespacial
Tiempo de permanencia de los datos	Breve (milésimas de seg)	Breve (7 a 40 seg)
Ejemplos	Percibir un sonido en medio de una multitud	Recordar la matrícula de un auto que acaba de pasar

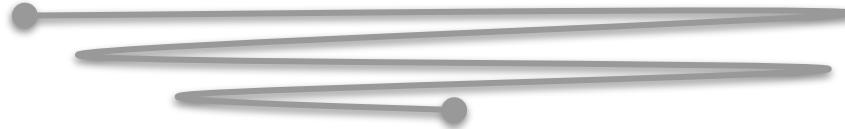
Fuente: <https://www.diferenciador.com/tipos-de-memoria/>

Tips para mejorar la visualización de datos - La memoria

	Memoria sensorial	Memoria a corto plazo	Memoria a largo plazo
Definición	Procesos cerebrales que interpretan estímulos por períodos mucho más breves que la memoria de corto plazo	Procesos cerebrales encargados de interpretar los estímulos y conservar esa información durante un tiempo breve	Procesos cerebrales encargados de conservar información durante períodos prolongados
Sistemas derivados	- Memoria icónica - Memoria ecoica - Memoria olfativa - Memoria gustativa - Memoria háptica	- Sistema ejecutivo - Almacén episódico - Bucle fonológico - Agenda visoespacial	- Memoria implícita - Memoria explícita
Tiempo de permanencia de los datos	Breve (milésimas de seg)	Breve (7 a 40 seg)	Prolongado (minutos a décadas)
Ejemplos	Percibir un sonido en medio de una multitud	Recordar la matrícula de un auto que acaba de pasar	Recordar cómo manejar bicicleta

Fuente: <https://www.diferenciador.com/tipos-de-memoria/>

Tips para mejorar la visualización de datos - La memoria



Hagamos un experimento para ver los límites de la memoria a corto plazo

Consigna:

1. Van a ver 5 series de letras
2. Tienen que recordarlas en el orden correcto

C X W

¿Las recuerdan?

C X W

M N K T Y

؟

M N K T Y

R P J H B Z S

؟

R P J H B Z S

G B M P V Q F J D

؟

G B M P V Q F J D

E G Q W J P B R H K A

؟

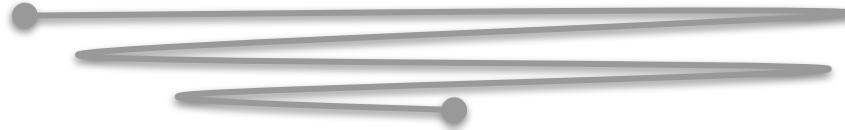
E G Q W J P B R H K A

Tips para mejorar la visualización de datos - La memoria



- En general (a la mayoría, después de sólo 1 lectura) las siguientes series les resultaron:
 - 1 -> extremadamente fácil
 - 2 -> fácil
 - 3 -> un poco + difícil
 - 4 -> sumamente difícil
 - 5 -> casi imposible
- Muestra la limitada capacidad de la memoria de corto plazo
- Se estima que en memoria a corto plazo se puede retener alrededor de 4 fragmentos de información visual
- Consecuencia. Resulta difícil recordar qué color representa cada categoría si se utilizan más de 4 colores/categorías diferentes en un gráfico de barras o columnas.

Tips para mejorar la visualización de datos



“Atributos Pre-atentivos” (Preattentive Attributes)

Tips para mejorar la visualización de datos - Preattentive attributes



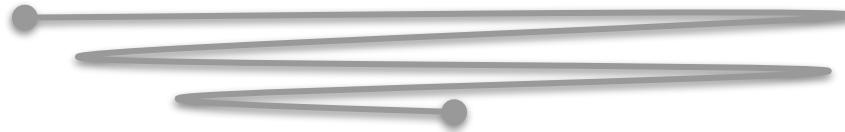
Atributos pre-atentivos representan aquellas características que pueden ser procesadas por la memoria icónica (memoria sensorial visual)

- Ver un gráfico o una tabla -> Ojos reciben estímulo y se transmite al cerebro
- Cerebro. Debe diferenciar y procesar

Percepción visual. Proceso mediante el cual nuestro cerebro interpreta esos reflejos producidos por la luz que entra por nuestros ojos. Relacionada con el funcionamiento de la memoria

- Memoria Icónica + Corto Plazo: las más importantes para el procesamiento visual
- Comprender qué aspectos se pueden procesar en estos tipos de memoria puede resultar útil para diseñar visualizaciones efectivas
- Ejemplo para ilustrar el poder de los atributos pre-atentivos

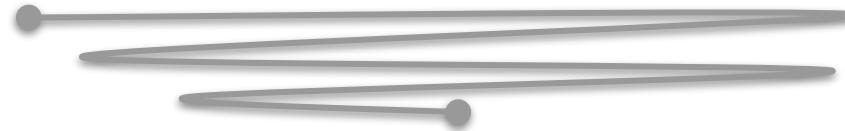
Tips para mejorar la visualización de datos - Preattentive attributes



Consigna ...

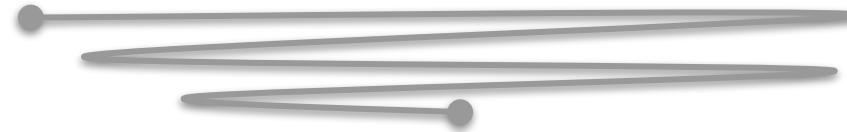
- Observar la siguiente figura
- Contar lo más rápido posible cuántas veces aparece el número de 7 en la figura

Tips para mejorar la visualización de datos - Preattentive attributes



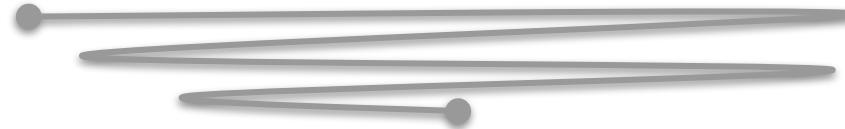
7	3	4	1	3	4	5	6	4	0
3	0	6	9	0	4	5	8	6	3
2	7	2	2	9	9	4	5	2	1
2	2	4	5	2	0	9	2	0	4
2	4	0	7	6	9	3	0	0	4
7	7	8	9	2	6	7	2	4	7
6	1	3	3	2	1	4	4	9	0
3	6	6	2	7	5	5	2	5	4
1	1	4	0	6	3	4	0	5	1
3	7	5	2	7	5	7	7	3	9
3	3	8	6	9	5	5	3	6	4
7	6	0	3	0	9	9	0	2	9
4	6	9	4	8	2	6	5	8	3
9	3	9	2	2	8	4	3	9	8
5	8	8	2	9	1	2	4	8	5
1	7	4	0	1	1	9	9	5	8

Tips para mejorar la visualización de datos - Preattentive attributes



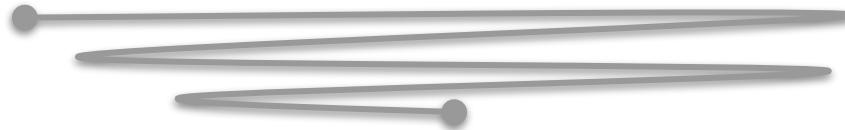
7	3	4	1	3	4	5	6	4	0
3	0	6	9	0	4	5	8	6	3
2	7	2	2	9	9	4	5	2	1
2	2	4	5	2	0	9	2	0	4
2	4	0	7	6	9	3	0	0	4
7	7	8	9	2	6	7	2	4	7
6	1	3	3	2	1	4	4	9	0
3	6	6	2	7	5	5	2	5	4
1	1	4	0	6	3	4	0	5	1
3	7	5	2	7	5	7	7	3	9
3	3	8	6	9	5	5	3	6	4
7	6	0	3	0	9	9	0	2	9
4	6	9	4	8	2	6	5	8	3
9	3	9	2	2	8	4	3	9	8
5	8	8	2	9	1	2	4	8	5
1	7	4	0	1	1	9	9	5	8

Tips para mejorar la visualización de datos - Preattentive attributes



7	3	4	1	3	4	5	6	4	0
3	0	6	9	0	4	5	8	6	3
2	7	2	2	9	9	4	5	2	1
2	2	4	5	2	0	9	2	0	4
2	4	0	7	6	9	3	0	0	4
7	7	8	9	2	6	7	2	4	7
6	1	3	3	2	1	4	4	9	0
3	6	6	2	7	5	5	2	5	4
1	1	4	0	6	3	4	0	5	1
3	7	5	2	7	5	7	7	3	9
3	3	8	6	9	5	5	3	6	4
7	6	0	3	0	9	9	0	2	9
4	6	9	4	8	2	6	5	8	3
9	3	9	2	2	8	4	3	9	8
5	8	8	2	9	1	2	4	8	5
1	7	4	0	1	1	9	9	5	8

Tips para mejorar la visualización de datos - Preattentive attributes



Reflexiones ...

- Respuesta correcta. "Hay 14 sietes"
- Figura 1. Lleva tiempo y es probable haber cometido un error
- Figuras 2 y 3. Uso de atributos pre-atentivos hace el trabajo más fácil
- Figura 2. Hace uso del color (naranja vs negro)
- Figura 3. Hace uso del tamaño (grande vs pequeño)
- El color y el tamaño se procesan en la memoria icónica y por eso podemos diferenciarlos tan rápidamente
- El uso adecuado de atributos pre-atentivos reduce la cantidad de esfuerzo necesario para procesar de manera precisa y eficiente la información que se comunica mediante una visualización de datos

Tips para mejorar la visualización de datos - Preattentive attributes



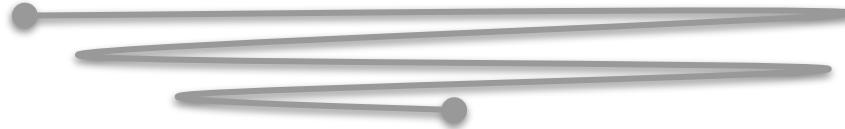
- *Preattentive attributes*
 - *Color*
 - *forma (incluye tamaño)*
 - *Ubicación*
 - *Movimiento*
- *Más info en el libro (ver bibliografia al final de la presentación)*

Cierre



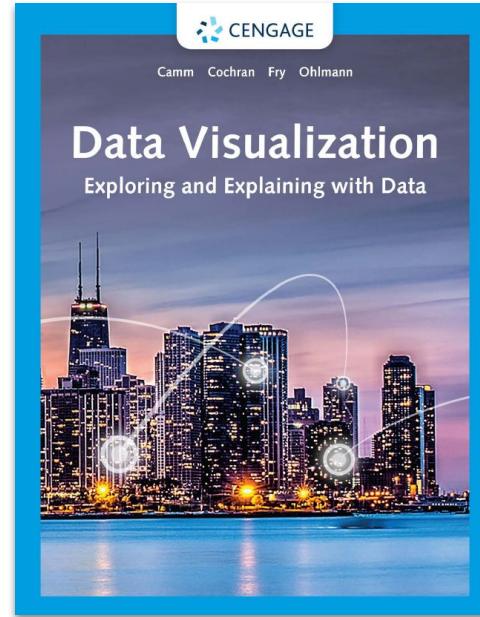
1. *Medidas de tendencia central y dispersión*
2. *Tips para mejorar la visualización*

Tareas para la próxima clase



1. *Resolver la guía de ejercicios de “Visualización”*

Bibliografia



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