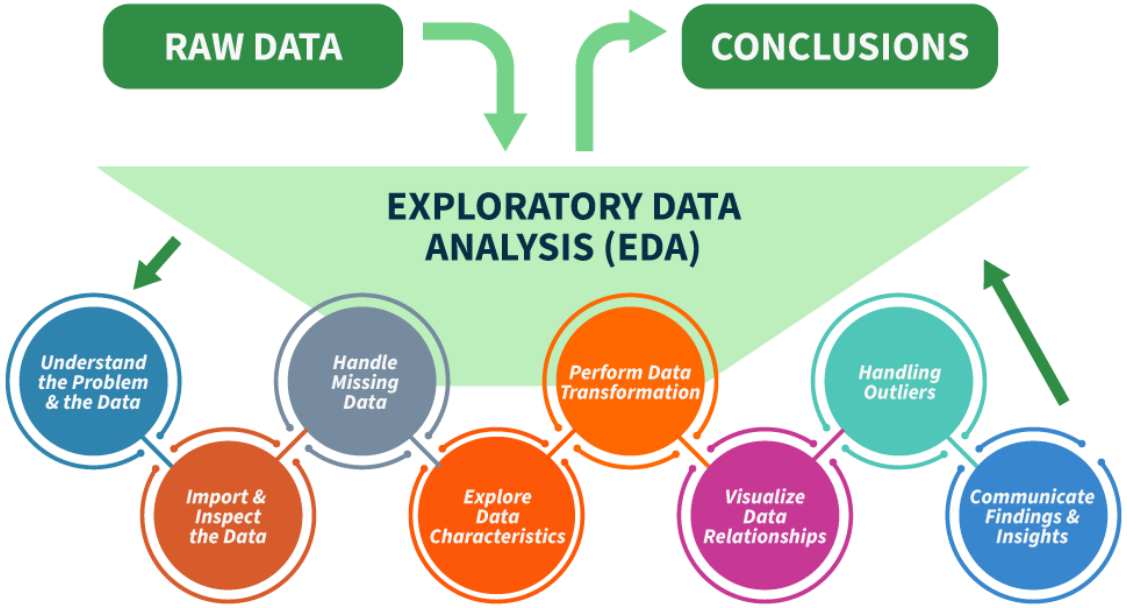
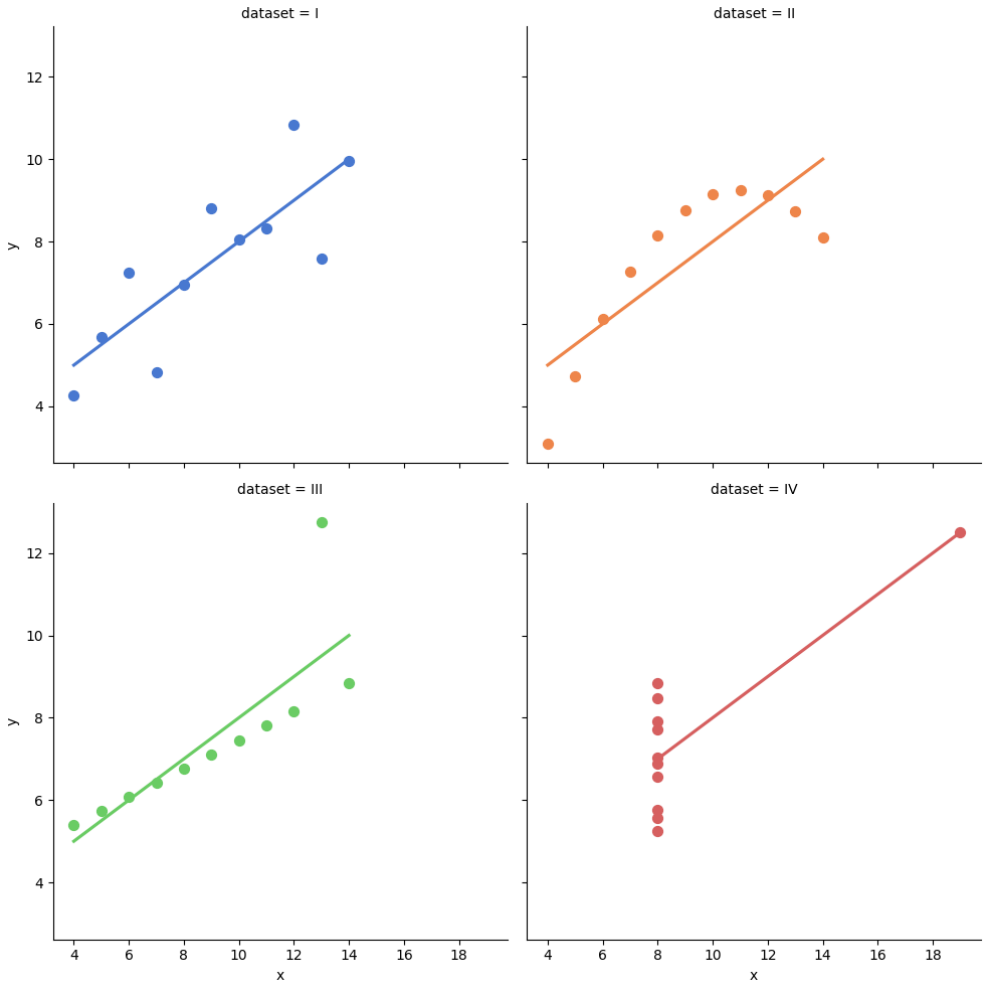
Análisis Exploratorio de los Datos



Profesor: Pedro Saa (pnsaa@uc.cl)
Año: 1-2025

¿Por qué necesito siempre hacer un análisis exploratorio de datos previo a un análisis estadístico?

El cuarteto de Ascombe nos enseña la importancia de primero visualizar los datos antes de comenzar un análisis estadístico



El análisis exploratorio de los datos nos permite:

- Entender el comportamiento de los datos a priori antes de nuestros supuestos
- Identificar errores obvios de experimentación
- Comprender mejor los patrones o relaciones que contien nuestros datos
- Identificar *outliers* o eventos anómalos

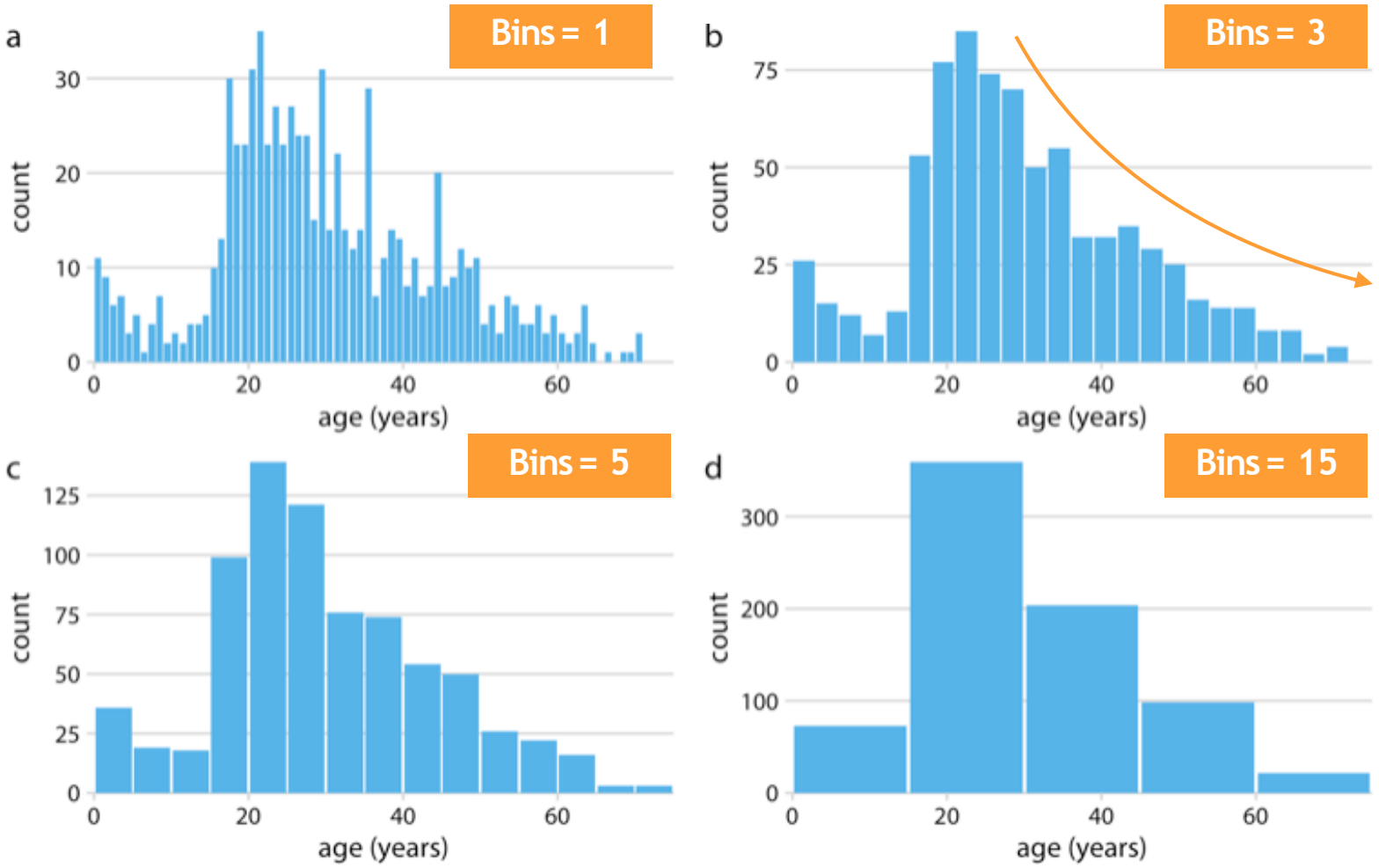
https://es.wikipedia.org/wiki/Cuarteto_de_Anscombe

F.J. Anscombe (1973) Graphs in Statistical Analysis," American Statistician, 27, 17-21.

Visualización de datos numéricos

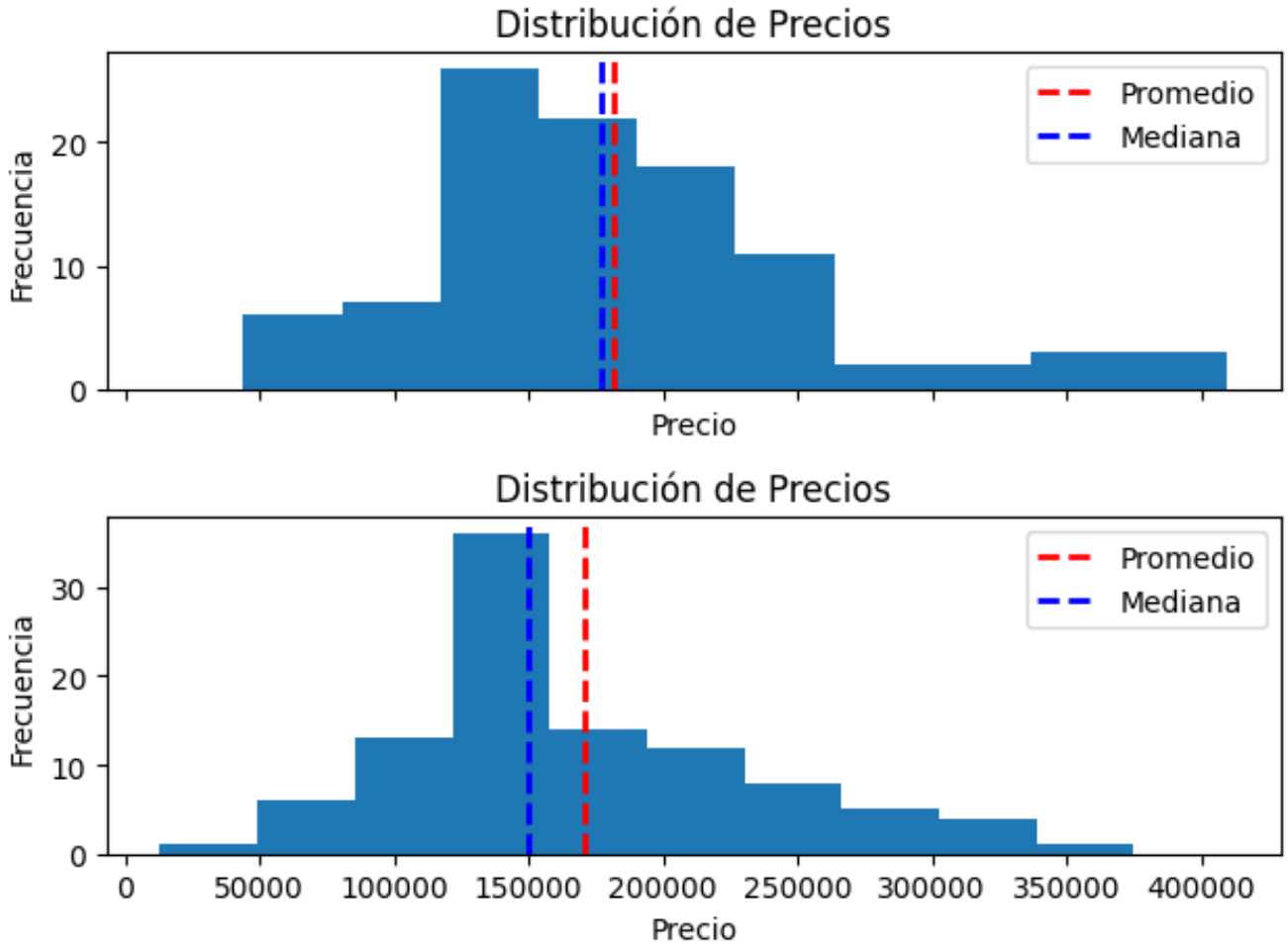
Los histogramas permiten visualizar la distribución o frecuencia (densidad) de una variable

Recomendable para bases medianas a grandes (*n*>20)

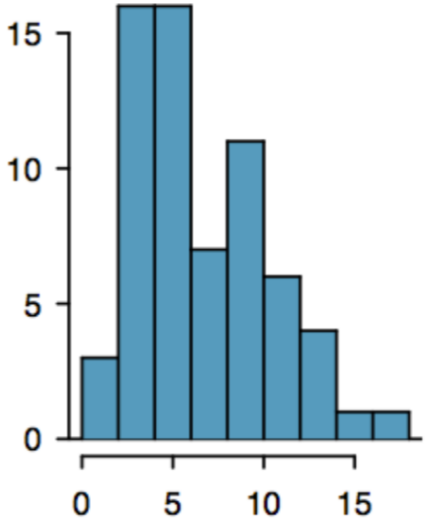


Los histogramas permiten visualizar la distribución o frecuencia (densidad) de una variable

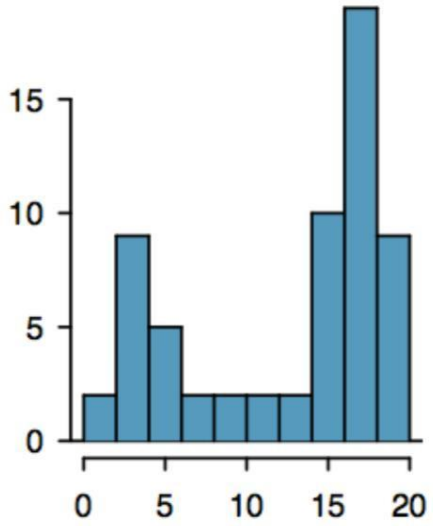
Recomendable para bases medianas a grandes (*n*>20)



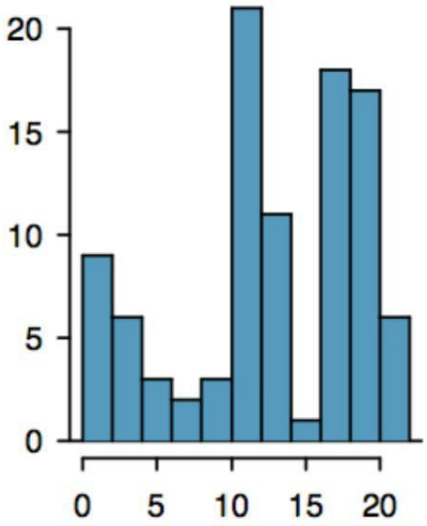
La distribución de una variable observada puede tener más de un *peak* o moda pudiendo relevar presencia de subgrupos



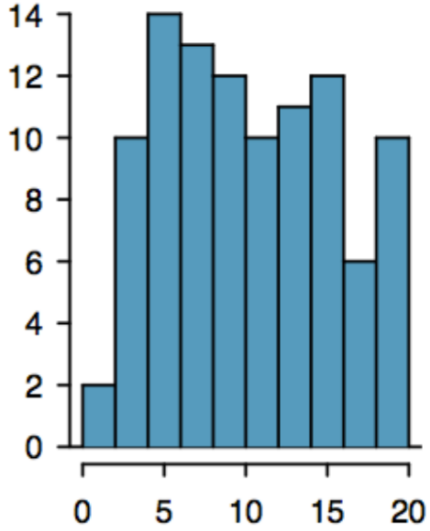
Unimodal



Bimodal



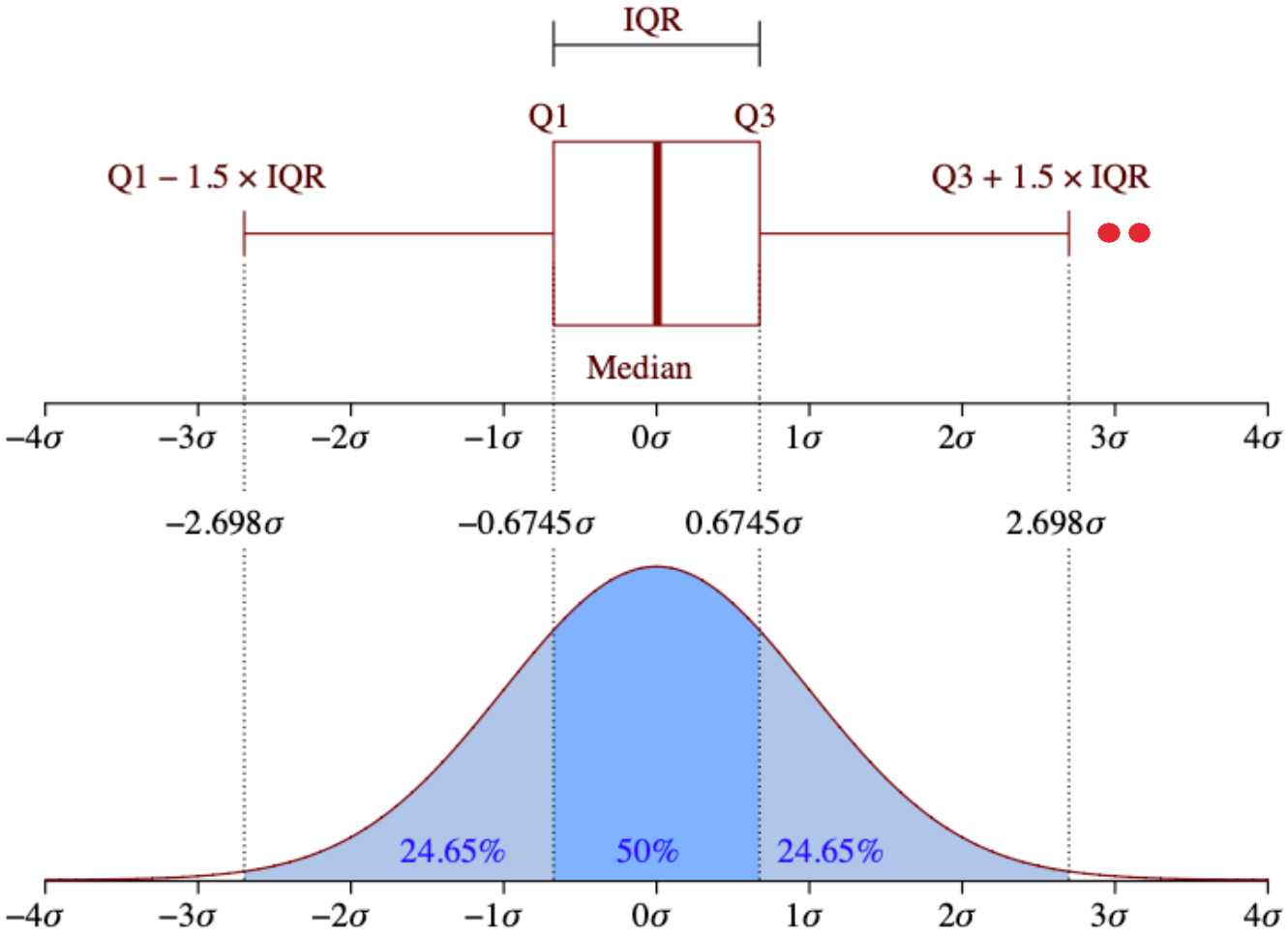
Multimodal



Uniforme

Los *boxplots* permiten visualizar información estadística relevante de una variable

Permite la visualización de valores atípicos o *outliers*



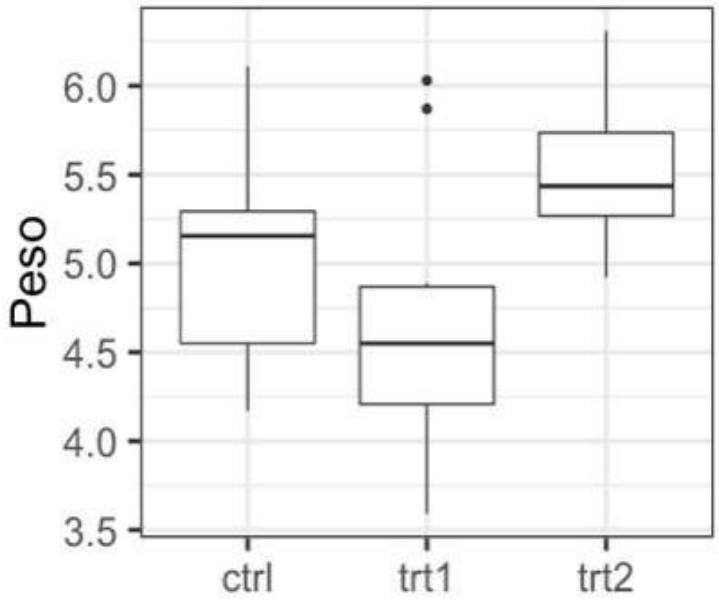
Boxplots informan sobre:

- Distribución de la variable
- Dispersión de la variable
- Tendencia central de la variable
- Valores atípicos

Los *boxplots* permiten visualizar información estadística relevante de una variable

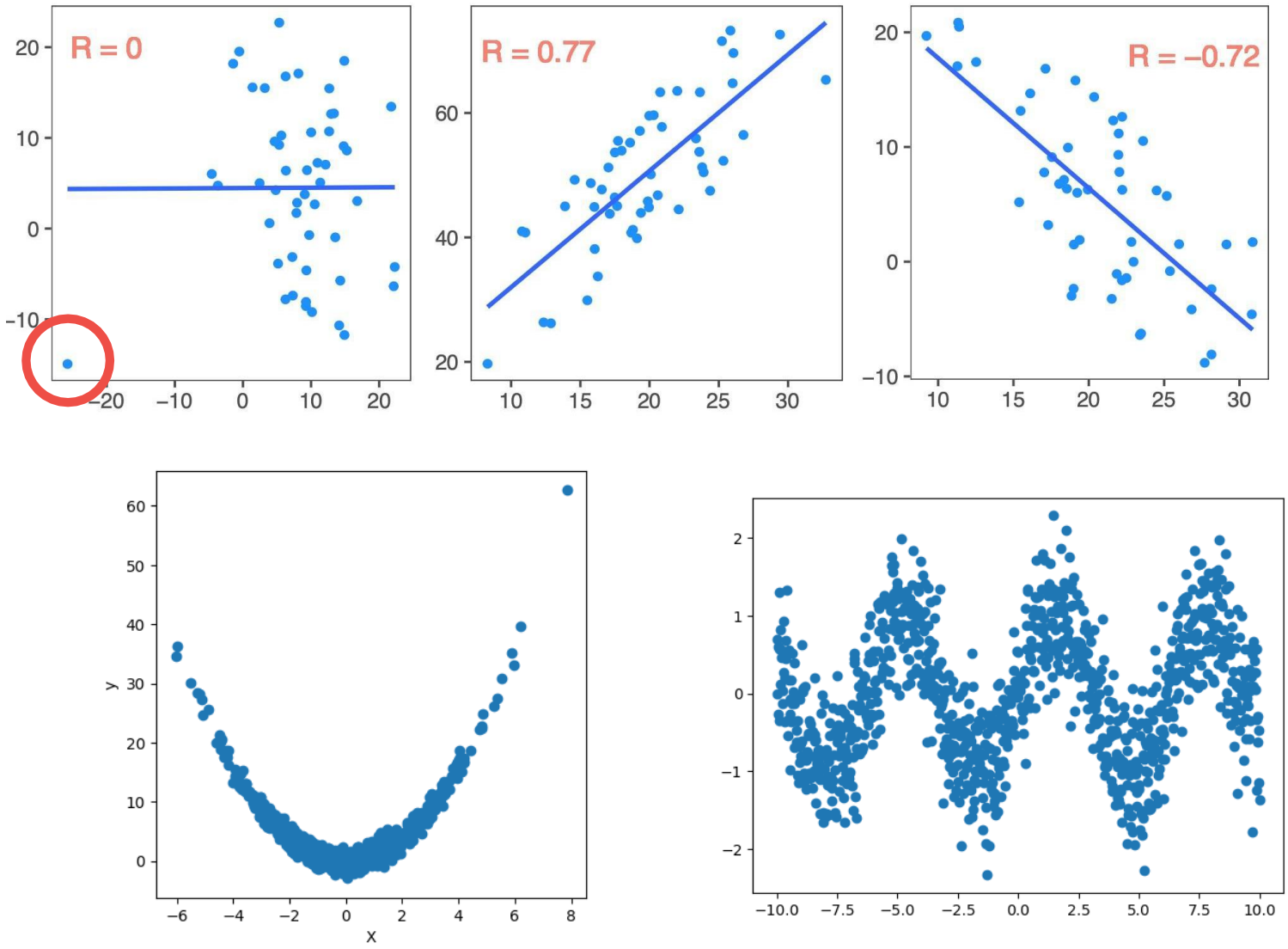
Permite la visualización de valores atípicos o *outliers*

Homocedasticidad



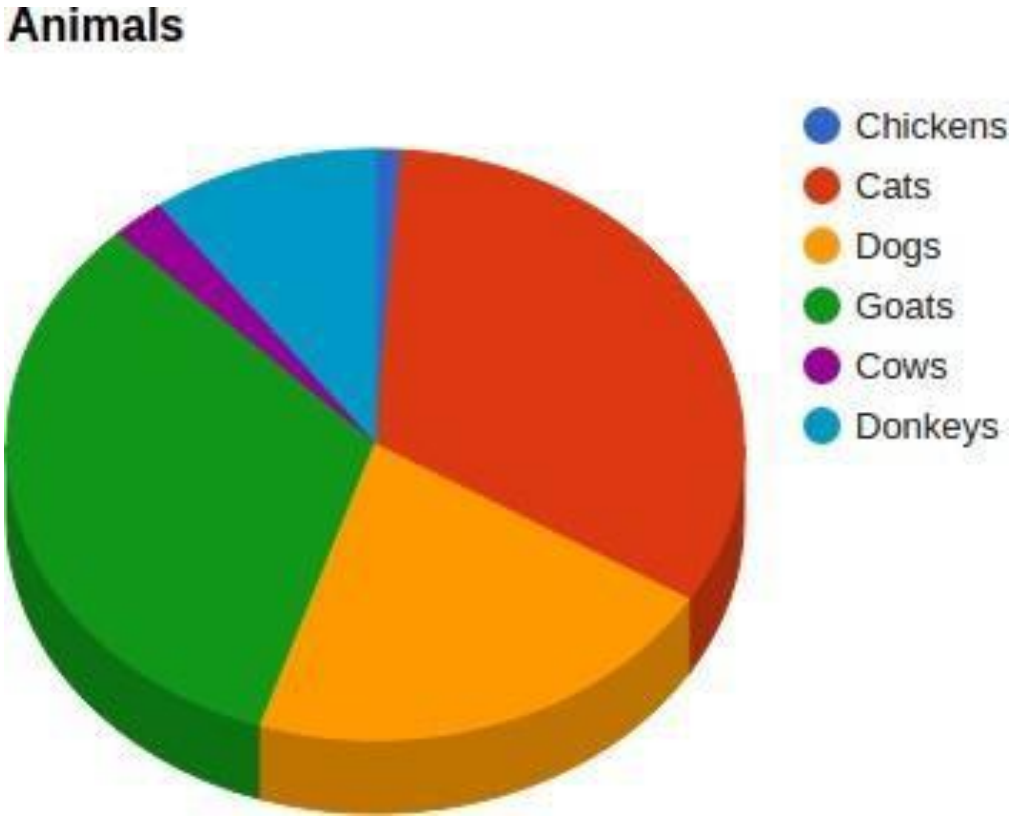
Observando boxplots, no debiese haber un grupo con una variabilidad visualmente muy distinta (ojalá no superior a 4 veces la SD de un grupo a otro)

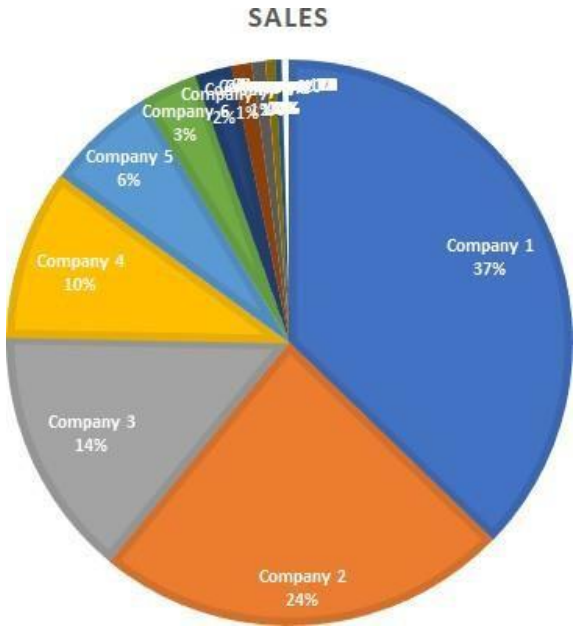
Los gráficos de dispersión permiten comparar dos variables numéricas e inspeccionar asociaciones y tendencias



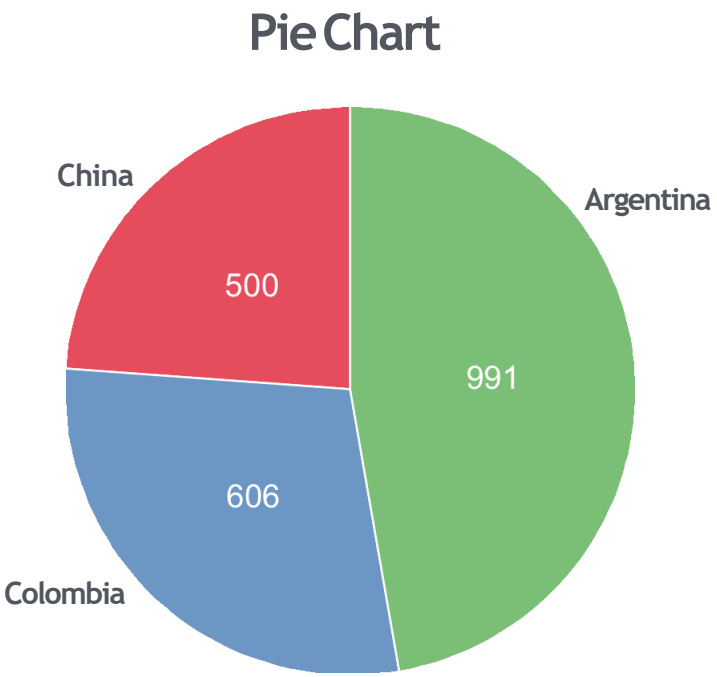
Visualización de datos categóricos

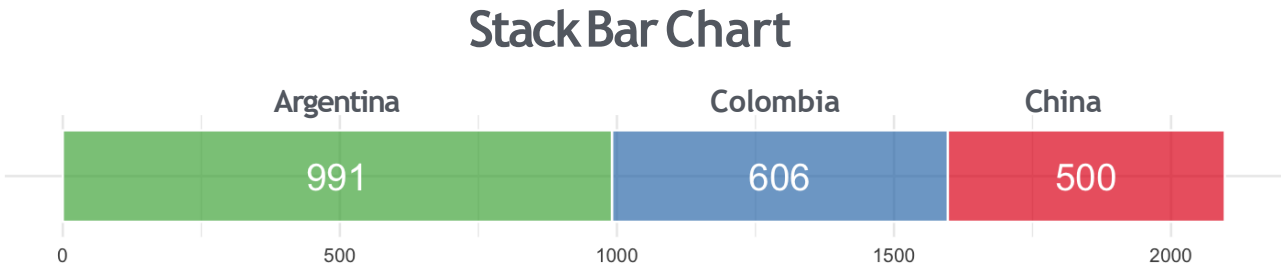
¿Opiniones?...

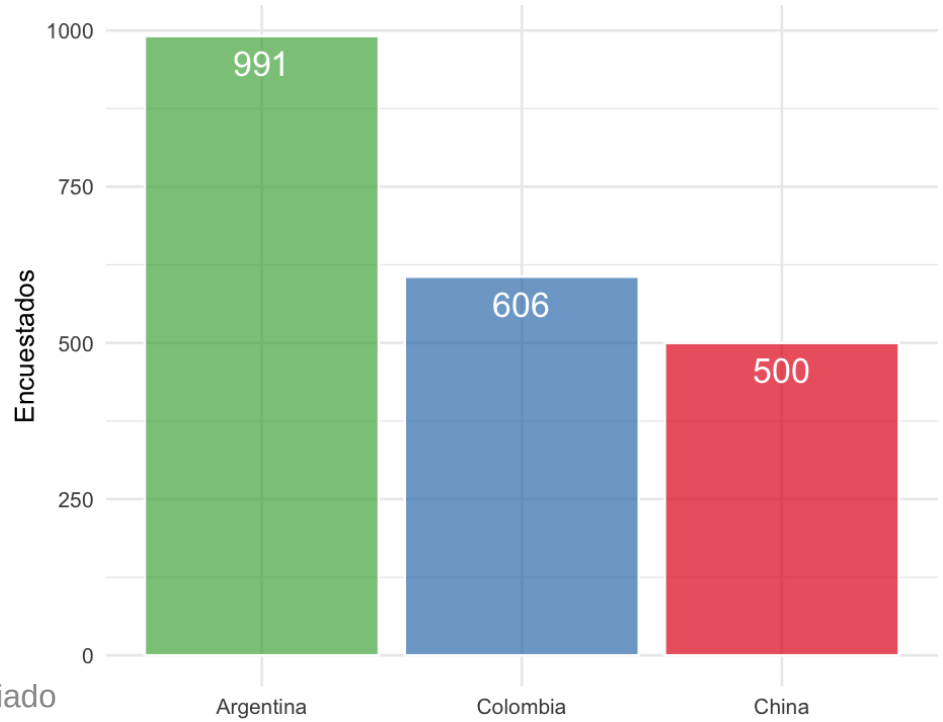

Animals
Chickens
Cats
Dogs
Goats
Cows
Donkeys


SALES
Company 1
37%
Company 2
24%
Company 3
14%
Company 4
10%
Company 5
6%
3%

Al ojo humano le cuesta cuantificar áreas
Lo mejor es simplificar al máximo la comparación



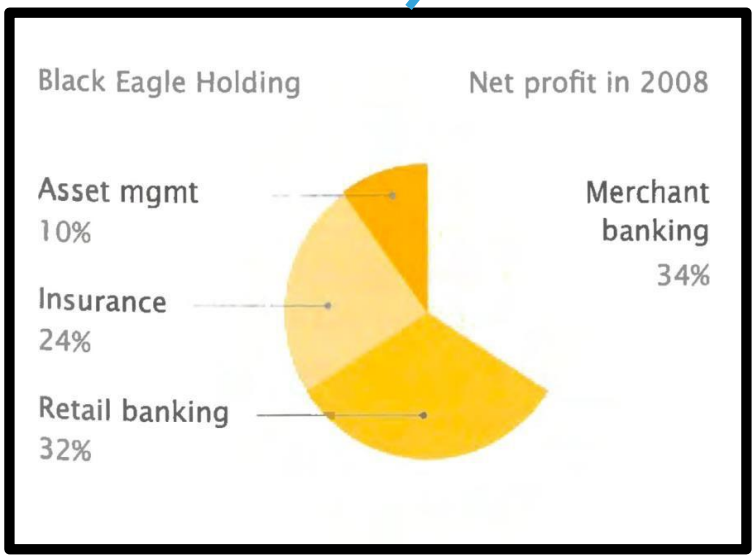


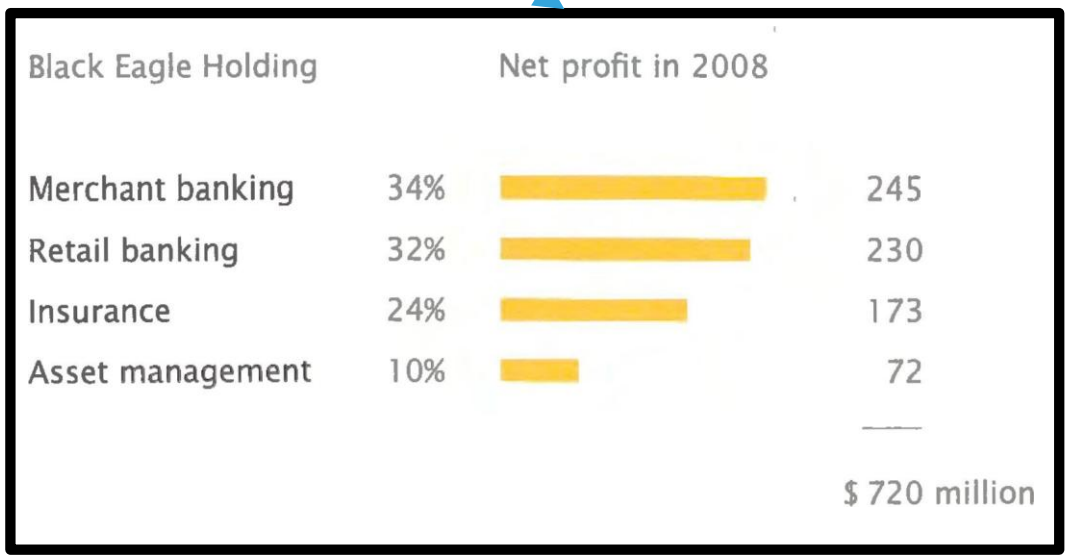


Datos: Encuesta de personas creyentes en países (Argentina, Colombia y China) el 2015

No es recomendable el uso de gráficos tortas, ya que nublan los datos y la interpretación es siempre mejor con un gráfico de barras

… Pero, ¿qué hago si tengo múltiples categorías relevantes?





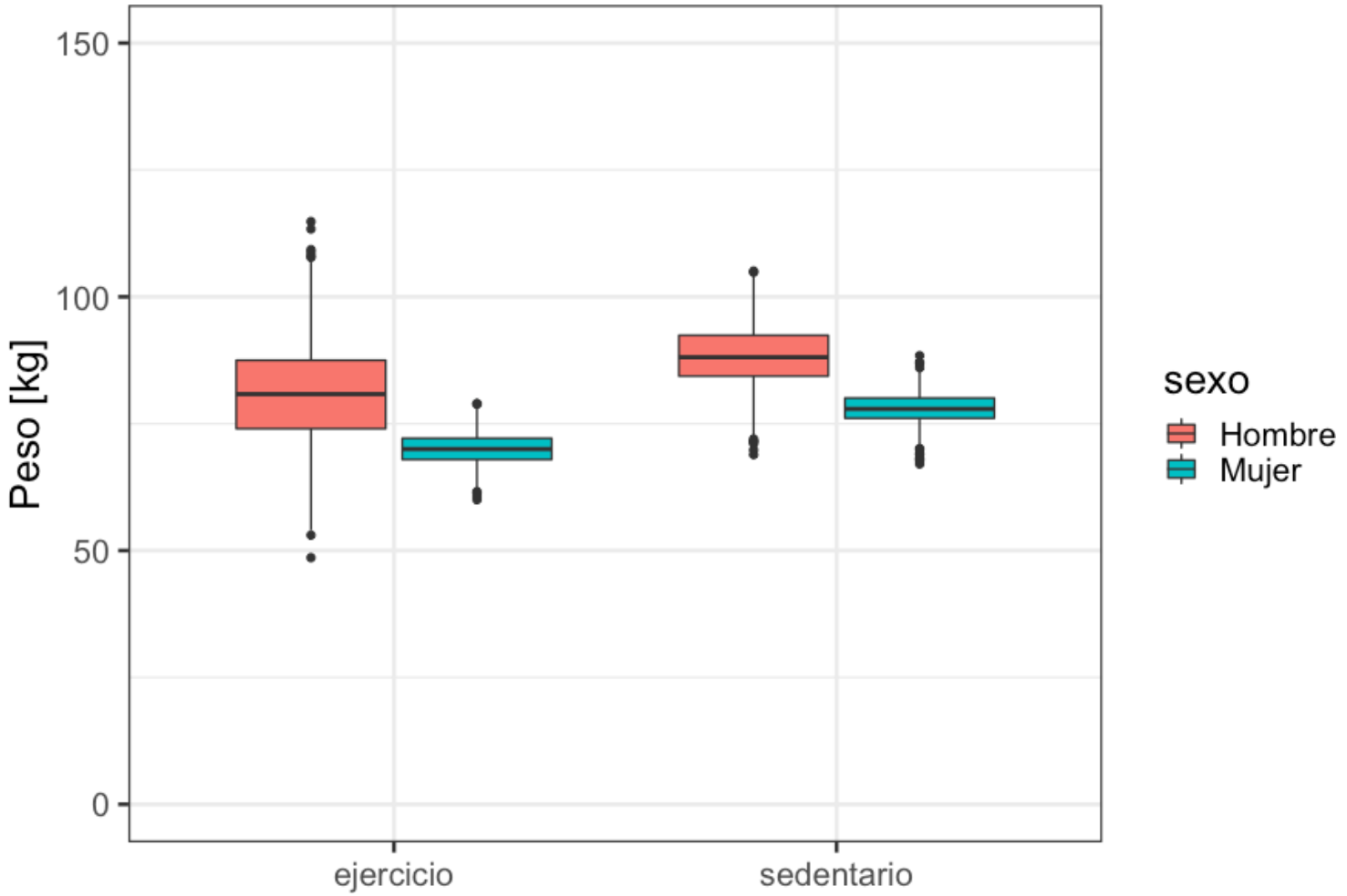
Depende de lo que se quiera **transmitir**

… ¿Queda más claro?

10% **De la rentabilidad ($72MM USD) lo provee Asset Management**

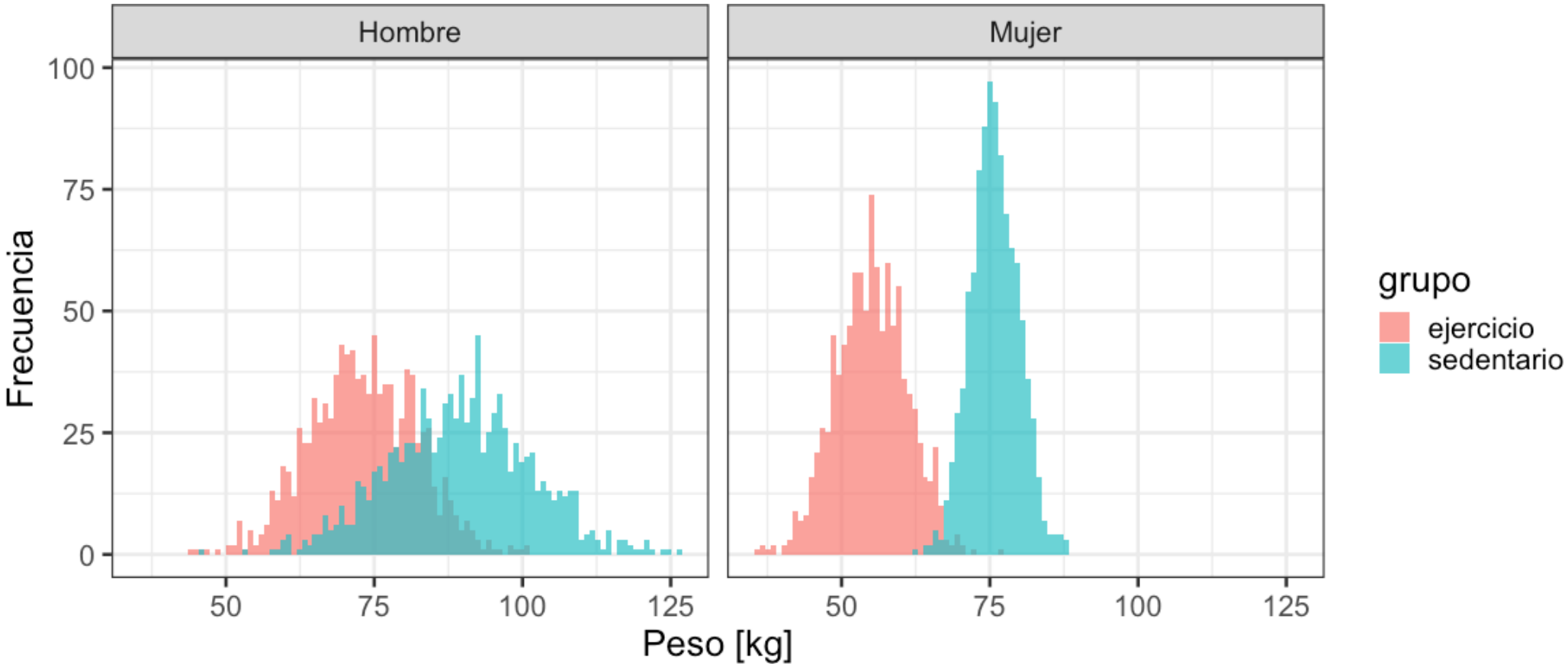
Los boxplots se pueden extender para variables categóricas

Permite comparar una variable numérica en distintas categorías



Los histogramas también se pueden extender para variables categóricas

Permite observar la distribución de distintos grupos de una variable



Las tablas de contingencias, si bien no son gráficos, permiten visualizar los datos categóricos resumidos de forma sencilla

Si bien la tabla anterior muestra resultados es difícil poder llegar a conclusiones, el siguiente resumen estadístico es bastante mejor:

	0-30 days		0-365 days	
	stroke	no event	stroke	no event
treatment	33	191	45	179
control	13	214	28	199
Total	46	405	73	378

Table 1.2: Descriptive statistics for the stent study.

20% de los pacientes sufrieron accidentes cerebro vasculares en el grupo de tratamiento al cabo de 1 año

12% de los pacientes sufrieron accidentes cerebro vasculares en el grupo control al cabo de 1 año

↑ **8%** Se observó este aumento de accidentes cerebro vasculares en el grupo de tratamiento vs control

Clase 0: Introducción al curso

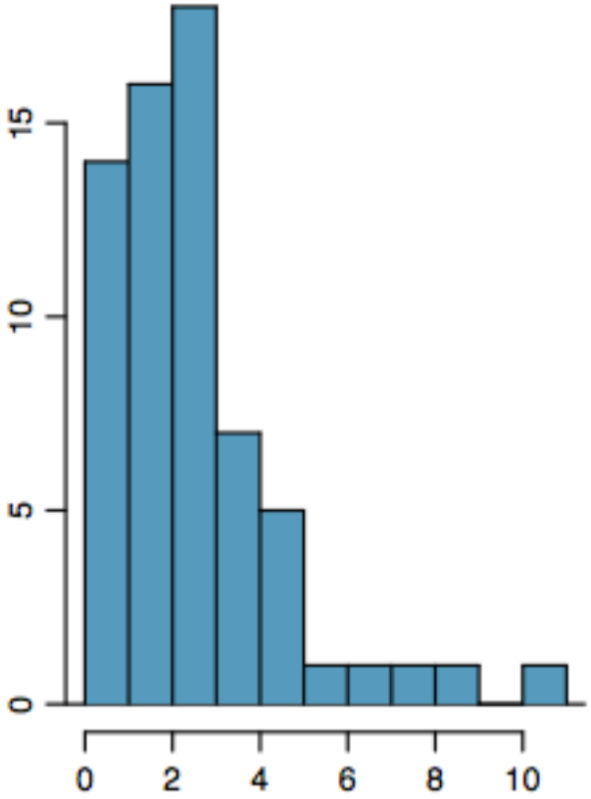
La clave es ver las proporciones entre las categorías

Cada tipo de gráfico tiene sus ventajas, limitaciones, y aplicaciones

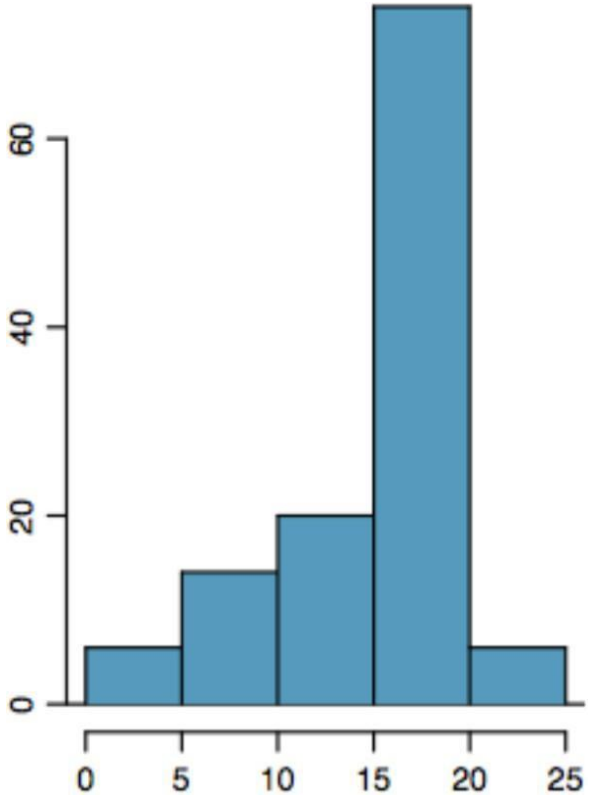
Tipo de gráfico	¿Cuándo ocupar el gráfico?
Dispersión (scatterplot)	Permite comparar dos variables numéricas, revisar asociaciones y tendencias.
Puntos (Dotplot)	Permite ver la distribución de una variable al apilar los puntos, recomendable para bases pequeñas ($n < 20$)
Histogramas	Visualizar la distribución o densidad de una variable, recomendable para bases medianas a grandes ($n > 20$)
Boxplots	Visualizar información estadística relevante de una o más variables (admite categóricas)
Barras	Visualizar variables categóricas (por ej. porcentajes, conteos)
Torta	Nunca*

Caracterización de datos categóricos

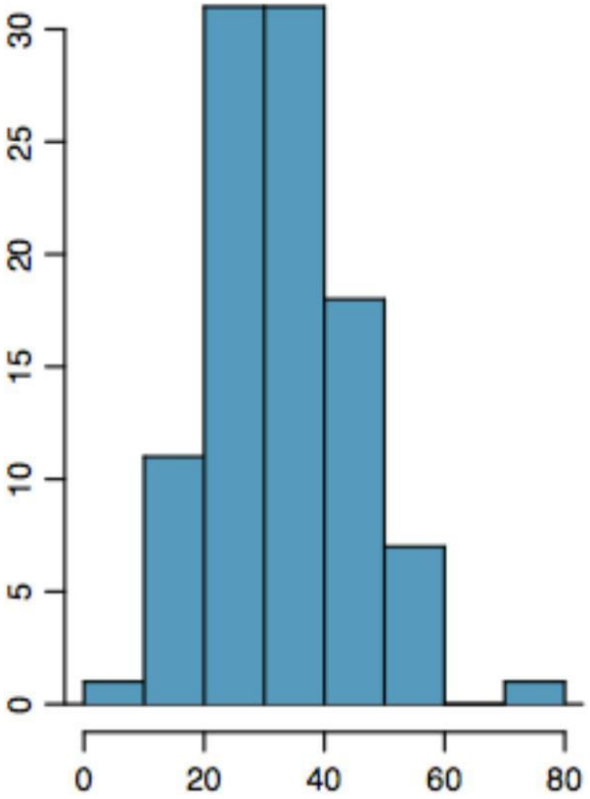
Una variable que tiene una distribución simétrica es generalmente considerada normal, no obstante, existen asimetrías cargada a la derecha o izquierda



Positiva o Derecha

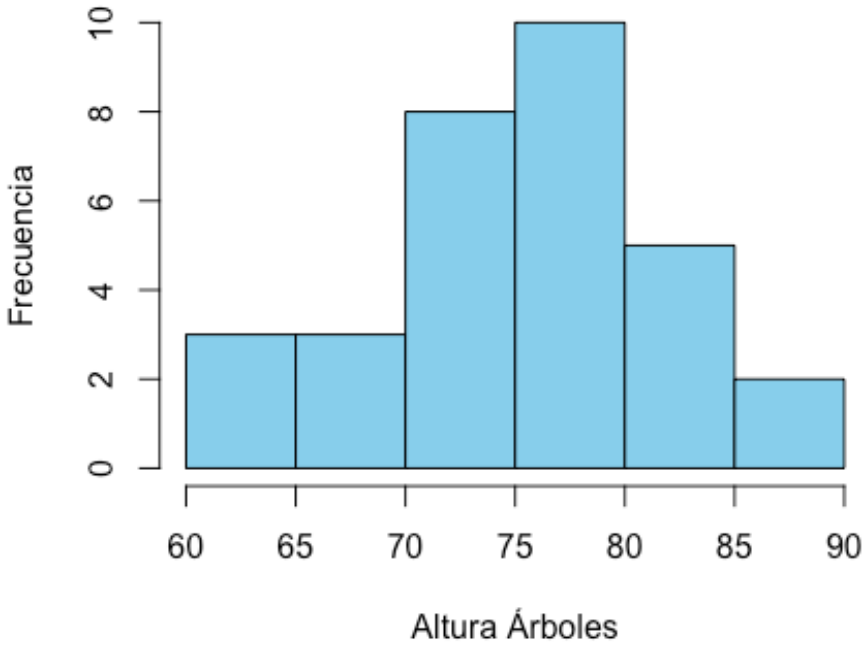


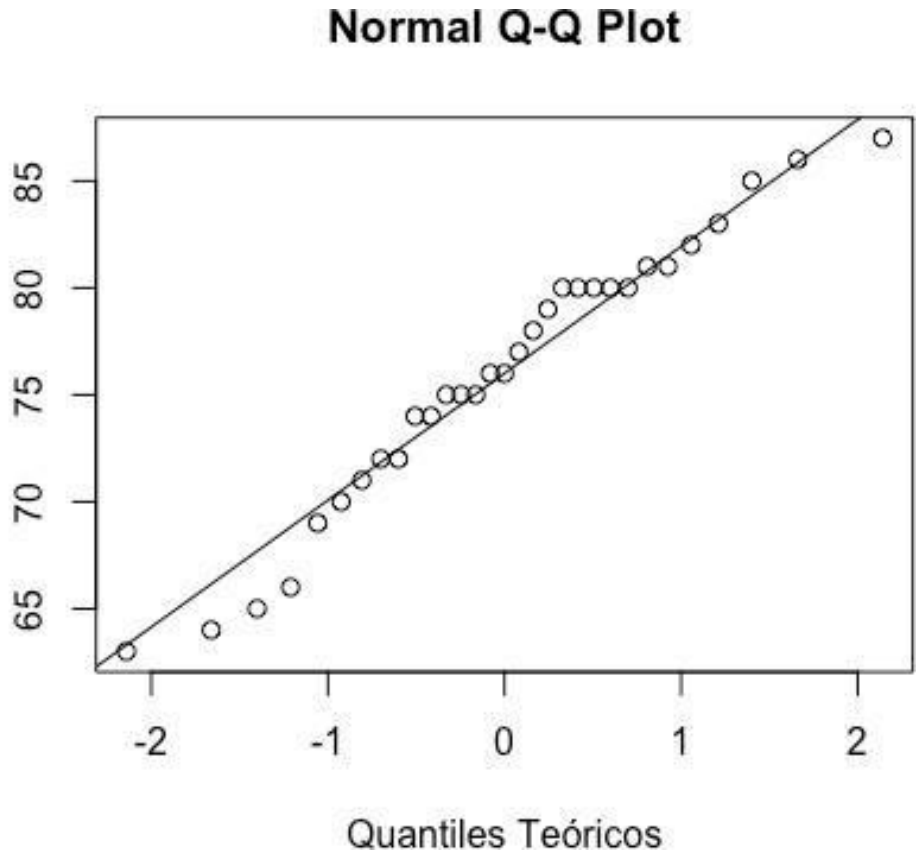
Negativa o Izquierda



Simétrica

Para evaluar la "normalidad" de una variable se puede hacer a través de un análisis visual usando los gráficos de probabilidad normal (**QQ-plot**)



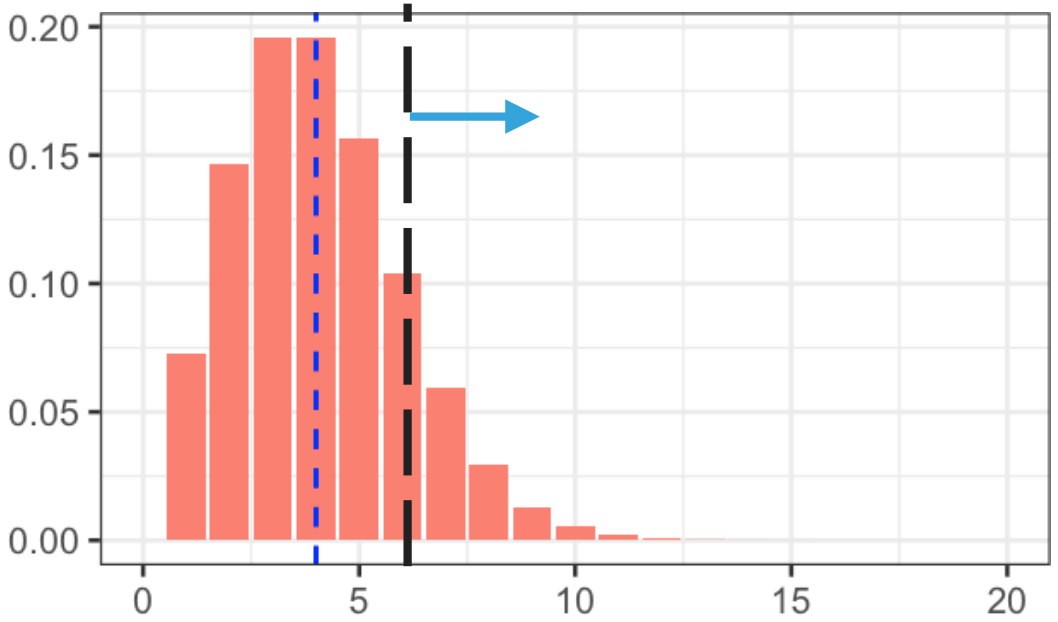


Mientras más se acerque los datos a la línea teórica diagonal más certeza tenemos que es normal

https://www.kaggle.com/code/gadaadhaarigeek/q-q-plot

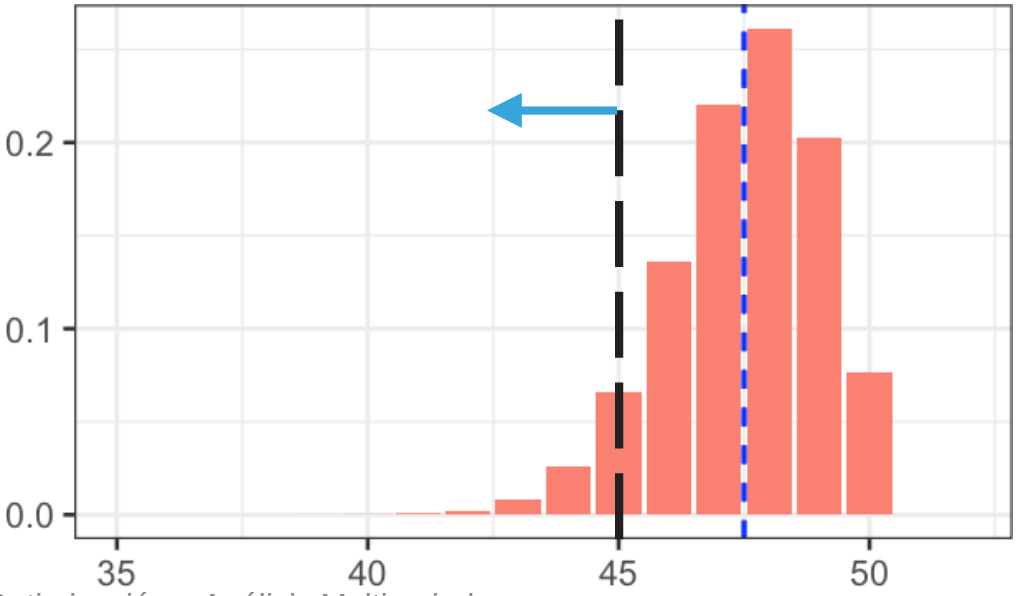
Para una distribución simétrica el promedio y mediana son muy similares, pero esto varía cuando encontramos asimetrías positivas o negativas

Promedio > Mediana



Promedio
Mediana

Promedio < Mediana



Existen tres (3) medidas de tendencia central; el promedio, la mediana y la moda, siendo las primeras dos las más importantes para el análisis

Puntajes examen

75, 69, 88, 93, 95, 54, 87, 88, 27

mean: $\frac{75+69+88+93+95+54+87+88+27}{9} = 75.11$

mode: 88

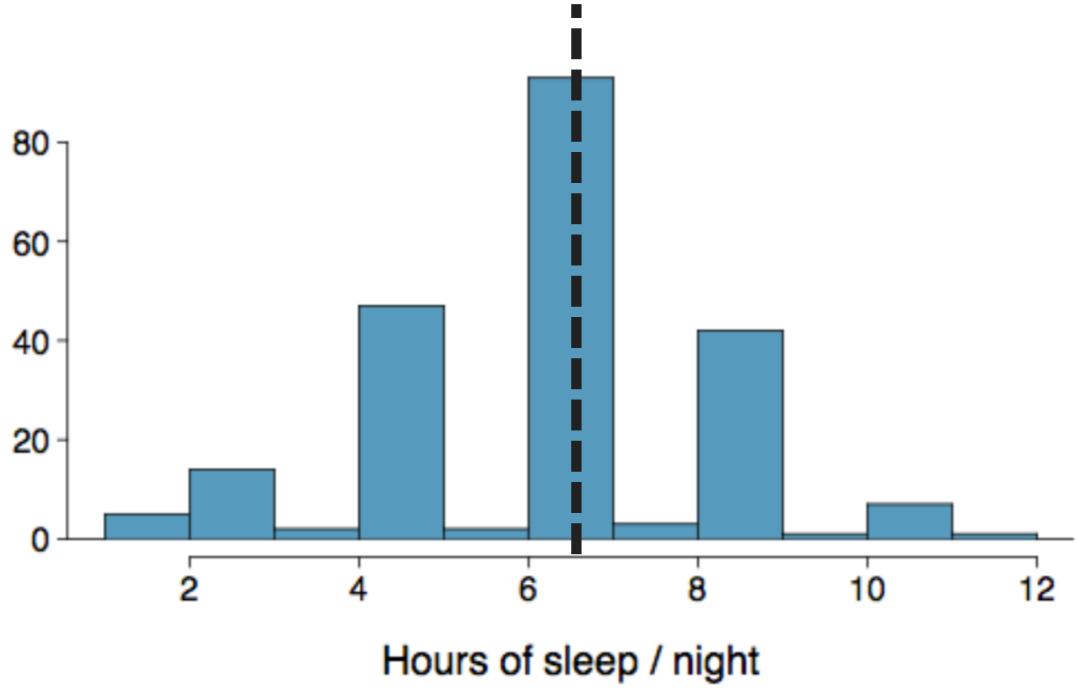
median: 27, 54, 69, 75, 87, 88, 88, 93, 95

Existen 3 medidas de dispersión relevantes: la primera es la varianza (s^2) o desviación cuadrada del promedio

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Sea la desviación = $(x_i - x_{prom})$
La varianza se eleva al cuadrado para eliminar desviaciones negativas y para pesar desviaciones grandes más fuertemente

Corrección de Bessel para estimación de valores poblacionales

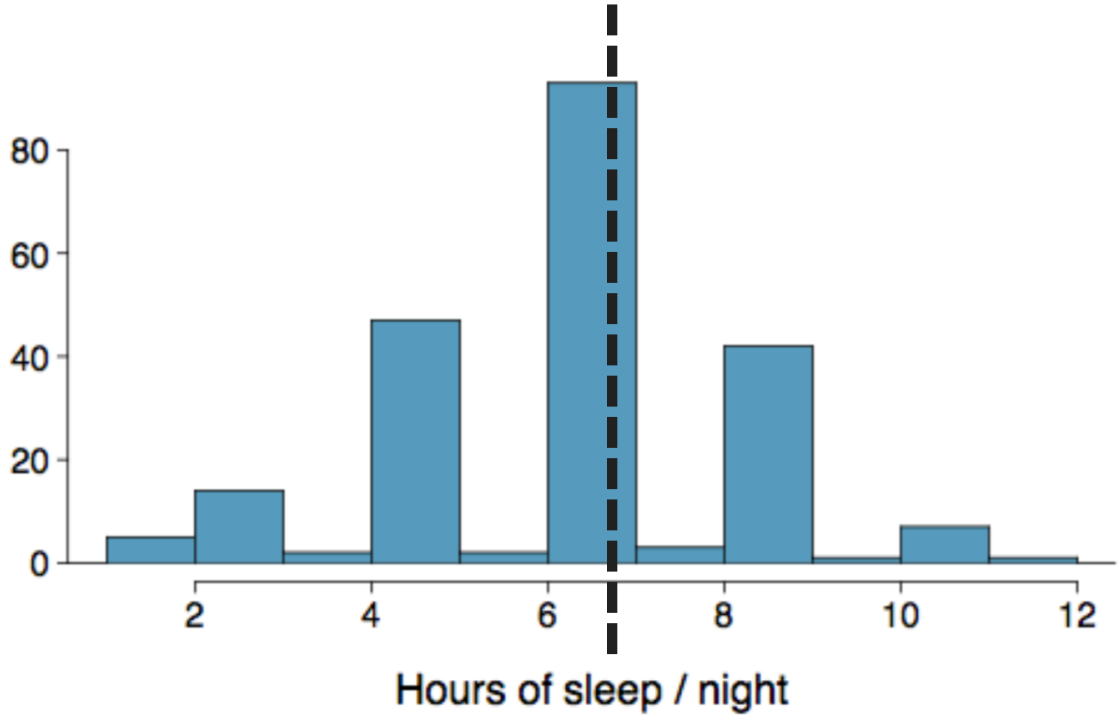


$$s^2 = \frac{(5-6.71)^2 + (9-6.71)^2 + \cdots + (7-6.71)^2}{217-1} = 4.11\ hours^2$$

Es difícil manejar las “unidades” de varianza, por lo que surge la idea de usar la desviación estándar (*s*)

$$s = \sqrt{s^2}$$

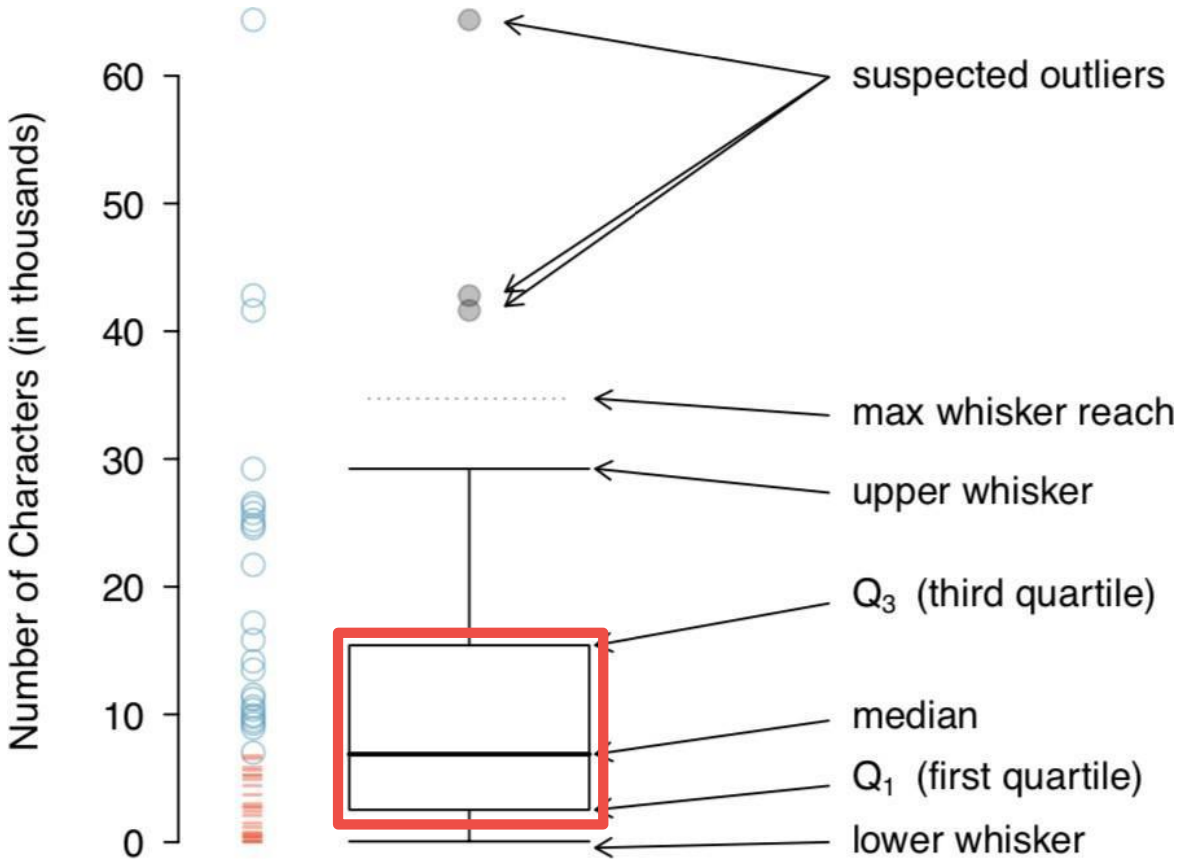
La desviación estándar tiene la misma unidad que los datos analizados



$$s^2 = \frac{(5-6.71)^2 + (9-6.71)^2 + \cdots + (7-6.71)^2}{217-1} = 4.11\ hours^2$$

$$s = \sqrt{4.11} = 2.03\ hours$$

El rango intercuartil (IQR), habitualmente observable en un boxplot, es el rango que contiene el 50% de los datos



Q1 = percentil 25 de los datos

Q2 = percentil 50 de los datos

Q3 = percentil 75 de los datos

IQR = Q3-Q1

Resumen

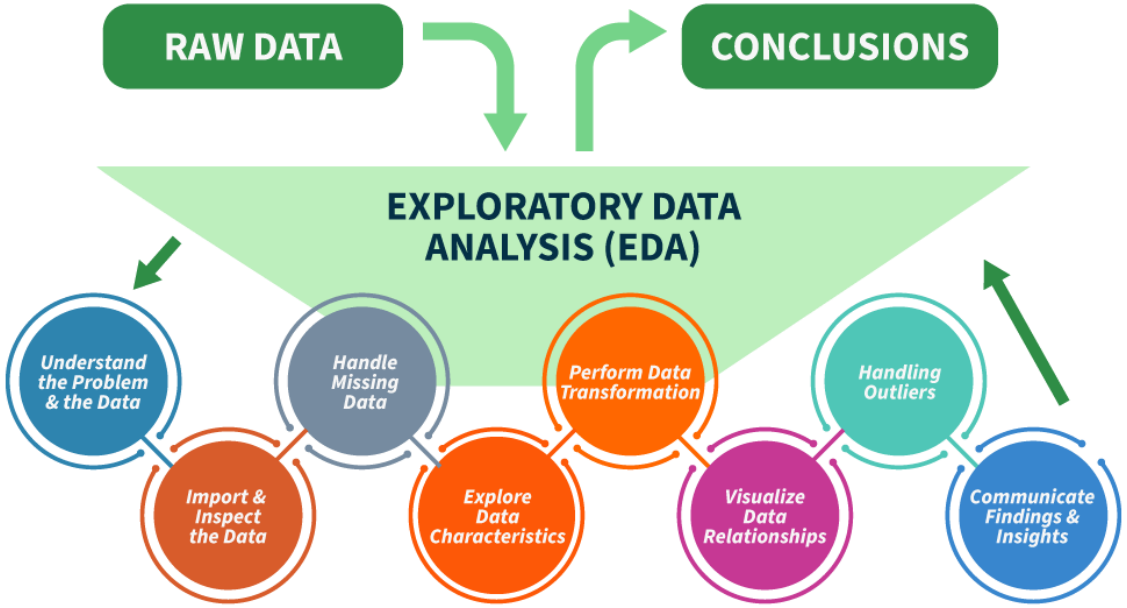
- El **Análisis Exploratorio de Datos** (EDA por sus siglas en inglés) es una herramienta muy útil para **entender mejor la estructura de los datos**, **identificar posibles problemas** con ellos, y **guiar los próximos análisis estadísticos.**

- Existen **diferentes estadísticos descriptivos o estadígrafos** que describen la estructura de los datos, ellos son de dos tipos: **tendencia central** o **dispersión**.

- **Diferentes tipos de gráficos pueden ser utilizados para visualizar el comportamiento de los datos.**

- La **combinación apropiada de gráficos con estadígrafos** permiten **comprender de forma más profunda el comportamiento de los datos.**

Análisis Exploratorio de los Datos



Profesor: Pedro Saa (pnsaa@uc.cl)
Año: 1-2025