TICS-411 Minería de Datos

Clase 2: Exploratory Data Analysis (EDA)

Alfonso Tobar-Arancibia
alfonso.tobar.a@edu.uai.cl

UNIVERSIDAD ADOLFO IBÁÑEZ

Análisis Exploratorio

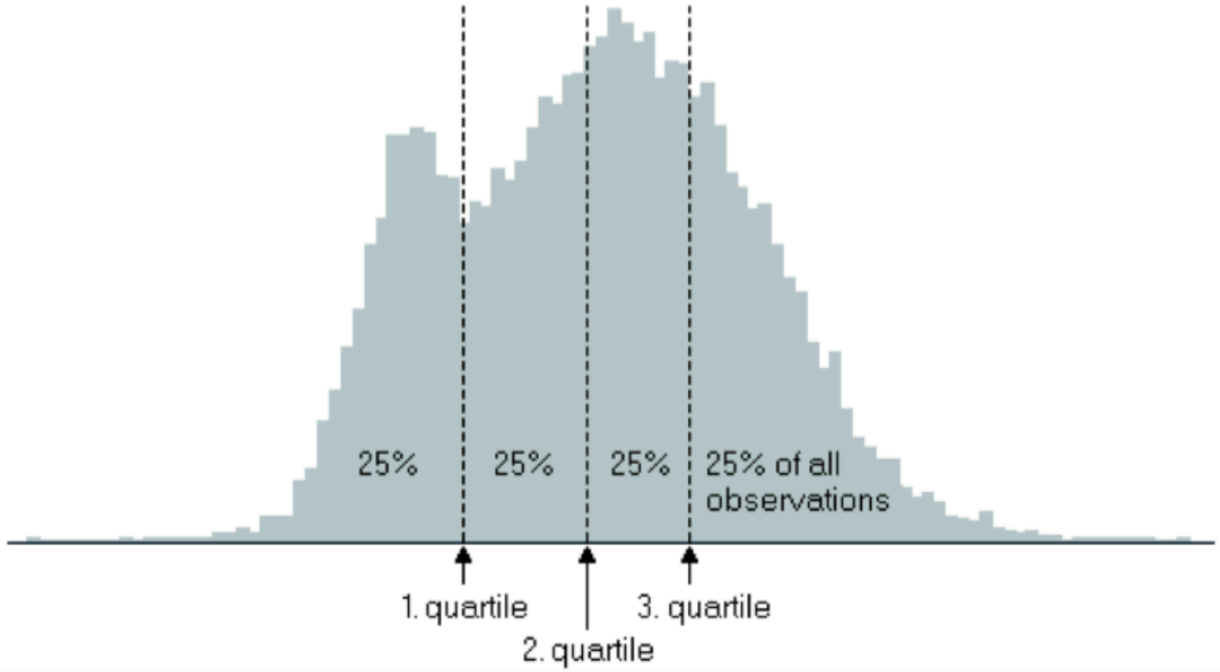
EDA

> El `Analisis Exploratorio de Datos` (EDA, por sus siglas en inglés) es procedimiento en el cual se analiza un dataset para explorar sus características principales.

EDA: Summary

Medidas de Tendencia Central

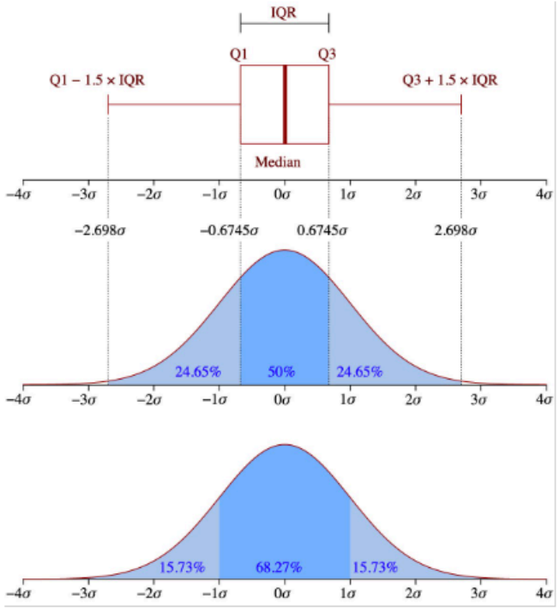
Mean	Mode	Median	Quartil
$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{m} x(i)$	Valor más repetido /popular de los datos	Valor con el 50% de los puntos arriba y 50% abajo	Valor con 25% (75%) de los puntos arriba y abajo

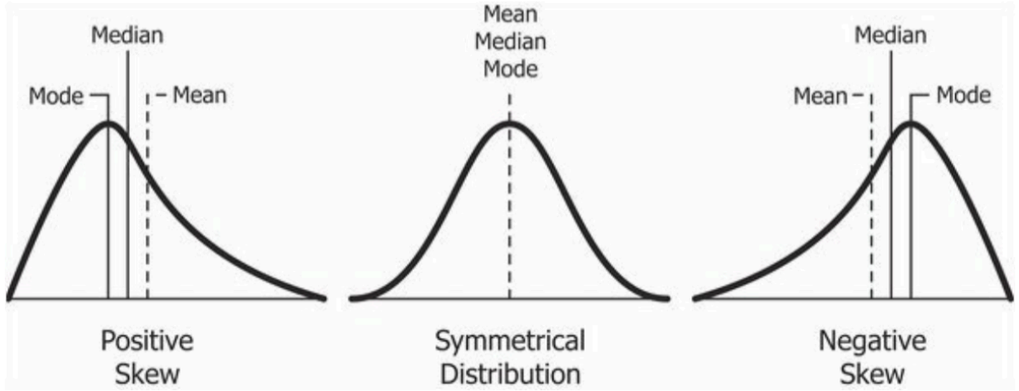


Medidas de Dispersión y Asimetría

Varianza	Rango	Rango Interquartil IQR
$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x(i) - \mu)^2$	Diferencia entre el maximo y mínimo valor	Diferencia entre el primero y tercer quaretil

Desviación Estándar	Skew
$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x(i) - \mu)^2}$	$\text{skewness} = \frac{\sum_{i=1}^{n}(x(i) - \mu)^3}{\left(\sum_{i=1}^{n}(x(i) - \mu)^3\right)^{\frac{2}{3}}}$





Visualizaciones

EDA: Visualización

> La visualización de datos es la presentación de datos en forma gráfica. Permite ver los análisis de modo que los responsables a cargo pueden comprender conceptos complejos.

> Gracias a la evoluación del cerebro humano somos capaces de detectar patrones complejos en la naturaleza a partir de la `Visión`.

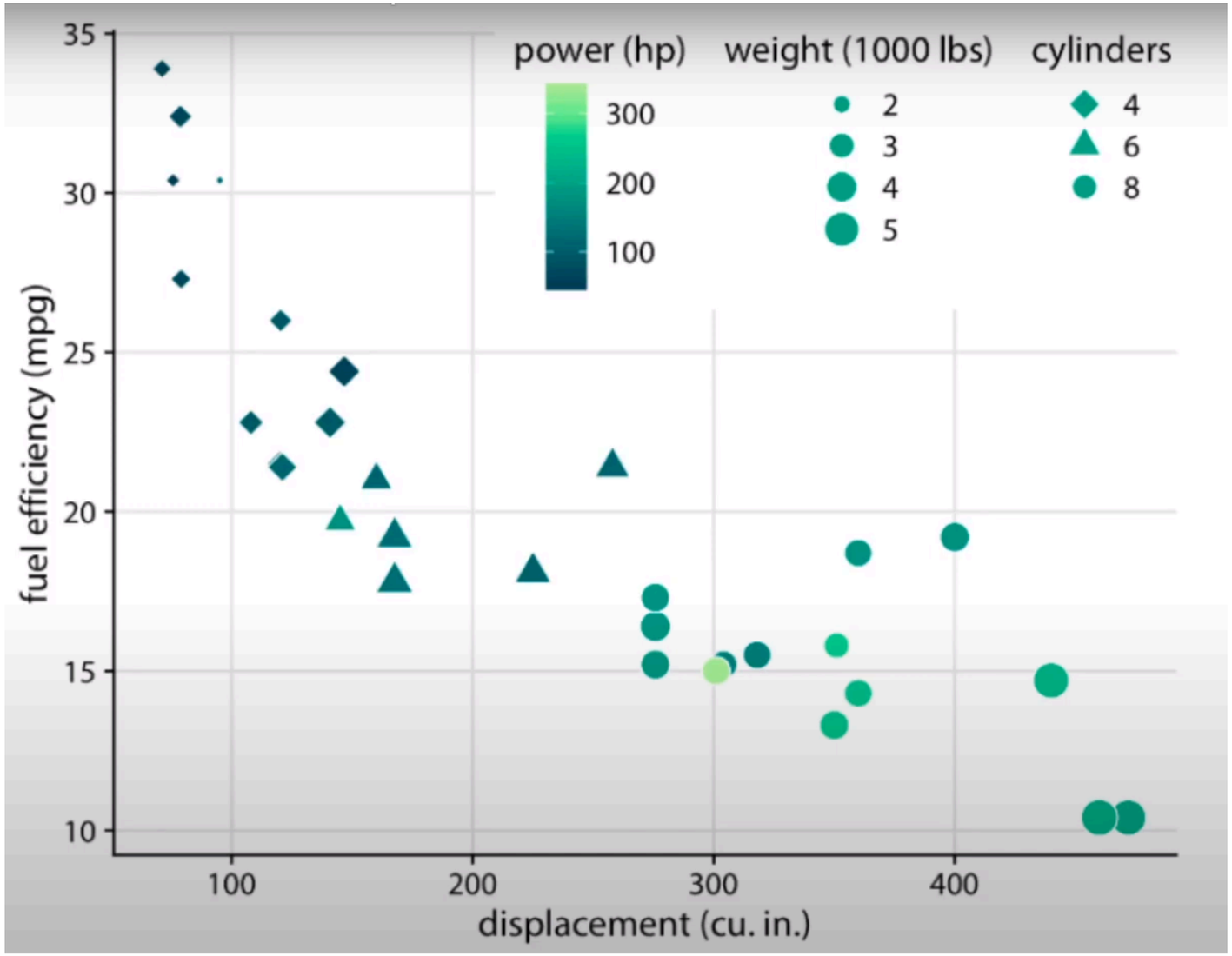
Puede ser difícil de aplicar si el tamaño de los datos es grande (sea en instancias o atributos). Por ejemplo, si los datos están en 4 dimensiones.

- Se suelen resumir los datos en `estadísticas simples`.
- Graficar datos en 1D, 2D y 3D (evitar dentro de lo posible).
- La visualización debe ser comprensible ojalá sin ninguna explicación.

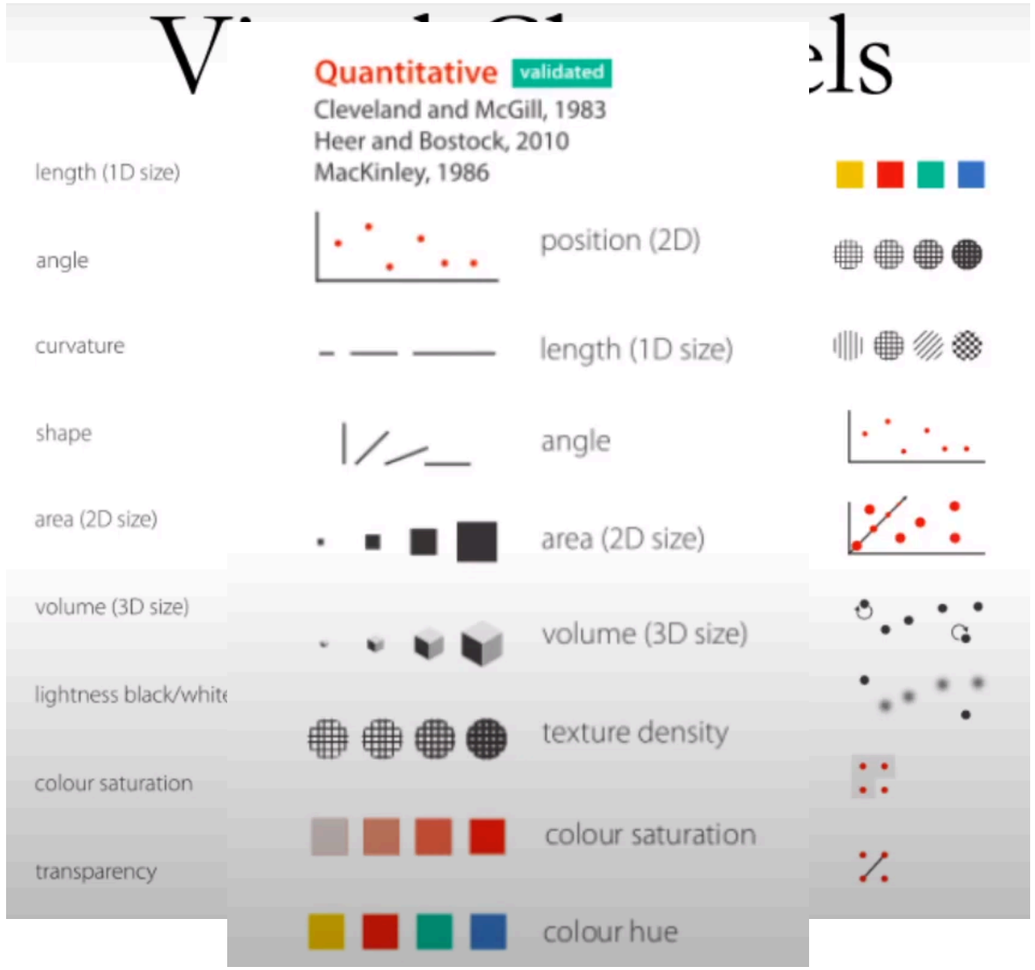
En caso de datos de alta dimensionalidad puede ser una buena idea reducir dimensiones mediante técnicas como:

- PCA
- UMAP
- etc.

Caso de Visualización



Canales Visuales



- Se les llama canales visuales a elementos visuales que pueden expresarse para poder expresar información **(Clase Visualizacion Andreas Mueller)**.
- La idea es poder mapear valores a cada uno de estos canales a valores.

- No todos los canales son igual de útiles ni fáciles de entender.

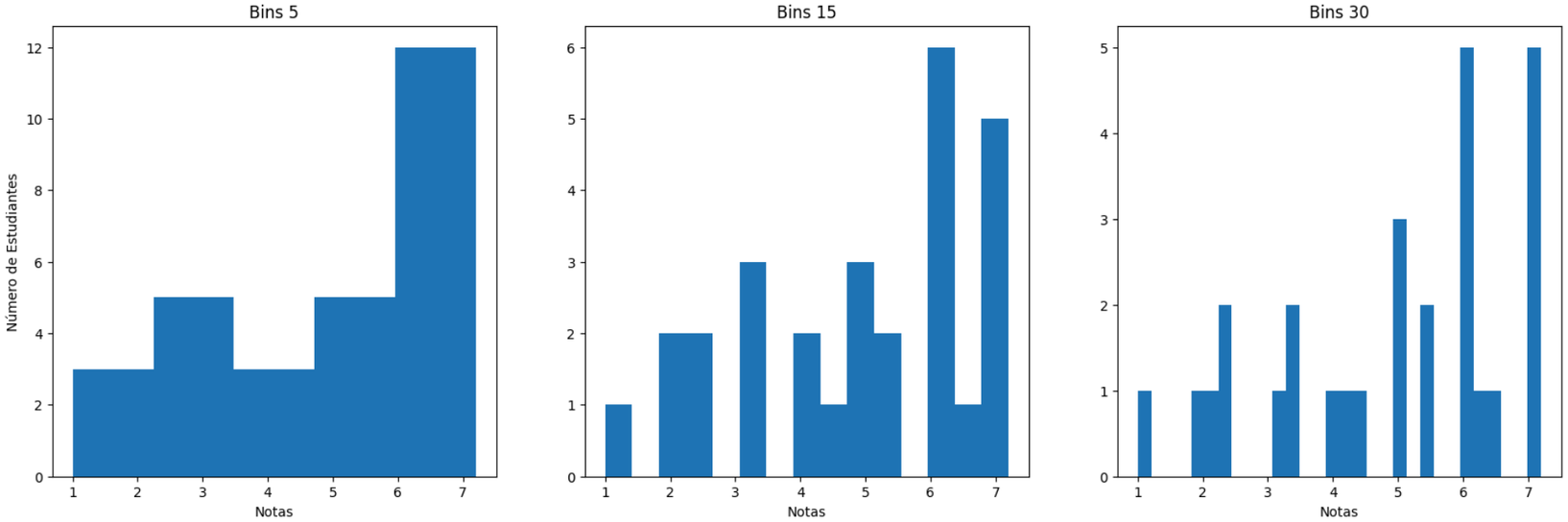
Visualizaciones: Distribuciones

Histograma

> El histograma permite visualizar distribuciones univariadas acumulando los datos en rangos de igual tamaño (`bins`).

- Permite visualizar el `centro`, la `extensión`, la `asimetría` y `outliers`.

- El histograma puede ser “engañoso” para conjuntos de datos pequeños.
- La visualización puede resultar de manera muy distintas dependiendo del número de `bins`.



Visualizaciones: Distribuciones

Kernel Density

> Corresponde a un suavizamiento de un Histograma en el cuál se usa un `Kernel` **(función no negativa que suma 1 y tiene media 0)** para agrupar los puntos vecinos.

La función estimada es:

$$f(x) = \frac{1}{n} = \sum_{i=1}^{n} K\left(\frac{x - x(i)}{h}\right)$$

- $K(u)$ es el Kernel.
- h es el ancho de banda.

- Gaussian kernel (`kernel = 'gaussian'`)

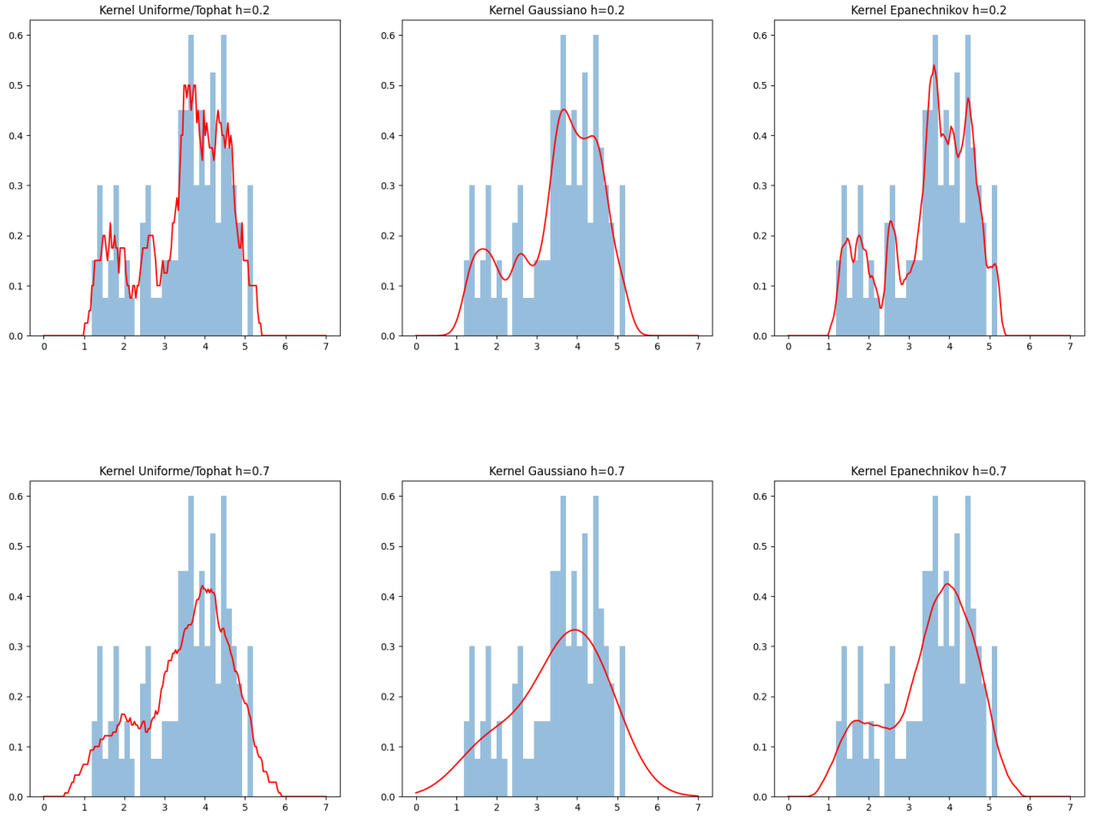
  $K(x; h) \propto \exp(-\frac{x^2}{2h^2})$

- Tophat kernel (`kernel = 'tophat'`)

  $K(x; h) \propto 1$ if $x < h$

- Epanechnikov kernel (`kernel = 'epanechnikov'`)

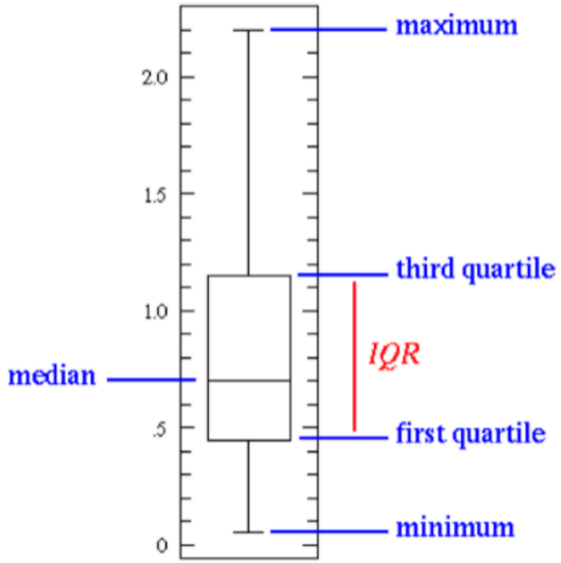
  $K(x; h) \propto 1 - \frac{x^2}{h^2}$

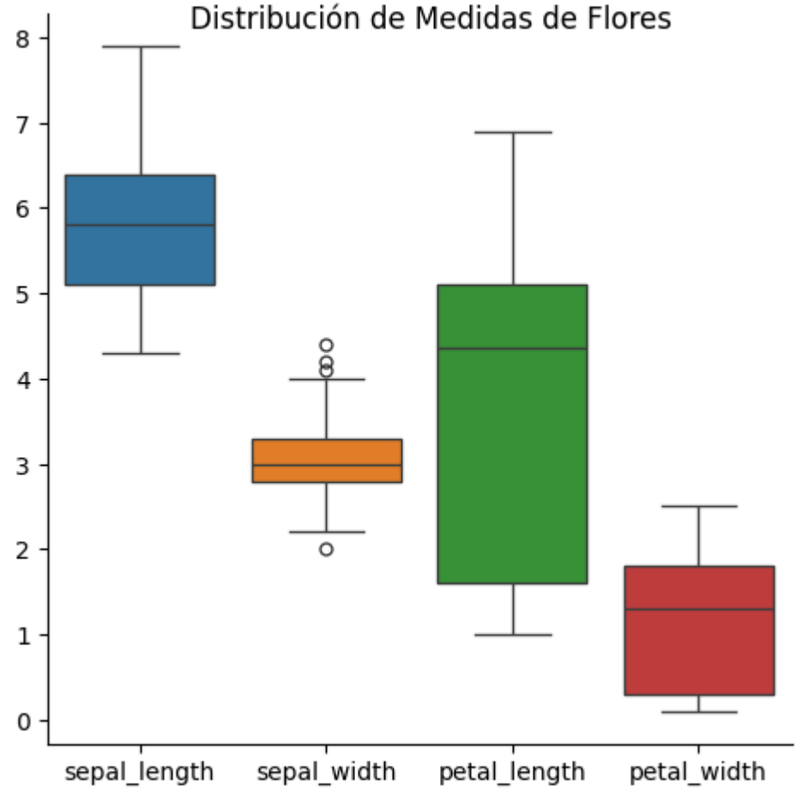


Visualizaciones: Distribuciones

Boxplot (Caja y Bigotes)

Es un tipo de gráfico que muestra la distribución de manera univariada.





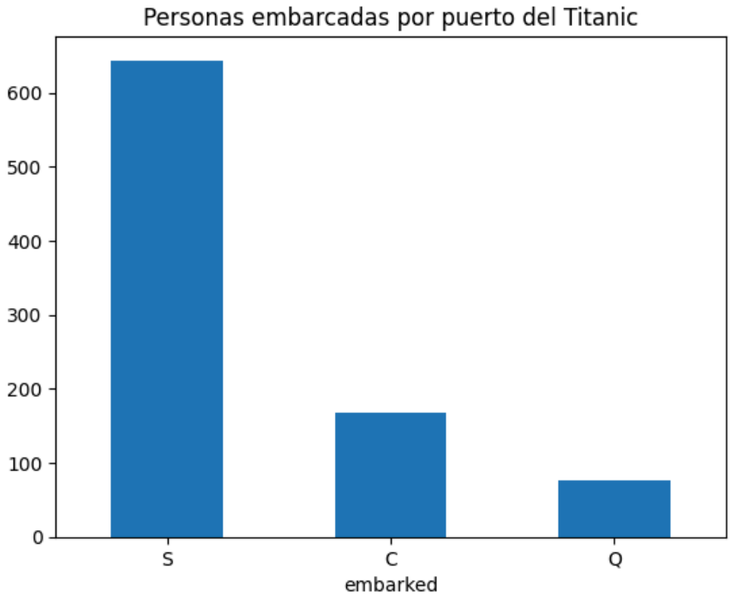
- Tiene la capacidad de mostrar varias distribuciones a la vez.
- Además presenta estadísticos de interés: Mediana, IQR y outliers.
- Los puntos fuera de los bigotes son considerados Outliers.

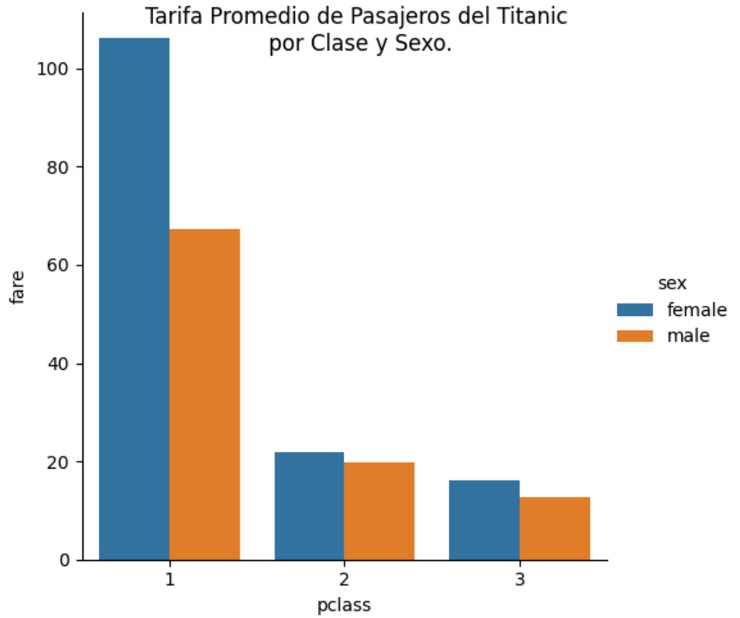
- Los bigotes pueden representar:
  - Mínimo y Máximo. (En este caso no hay outliers).
  - $\mu \pm 3\sigma$
  - Percentiles 5 y 95.
  - Otros valores.

Visualizaciones: Barras

Bar Plot

> La altura de la barra (normalmente Eje y) representa una cantidad asociada a una categoría (normalmente Eje x).





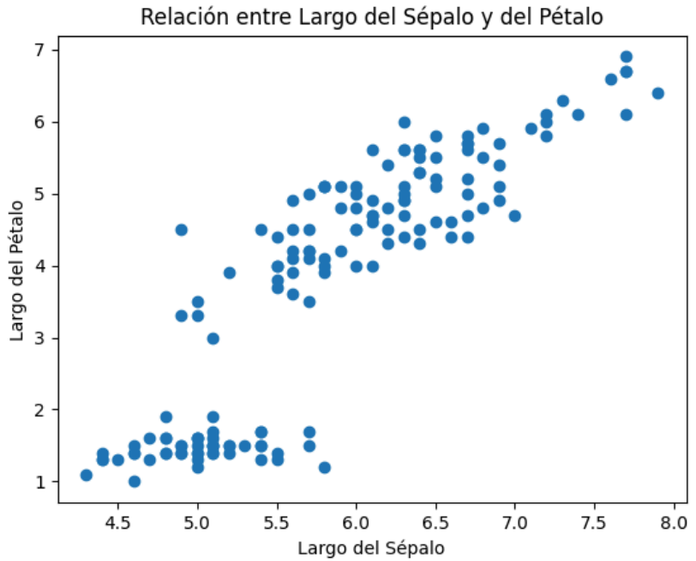
⚠ • Otras convenciones llaman a este gráfico `Column Plot`, mientras que el `Bar Plot` tiene las barras de manera horizontal.

Visualizaciones: Puntos

Scatter

> Gráfico empleado para mostrar distribución de datos bivariados

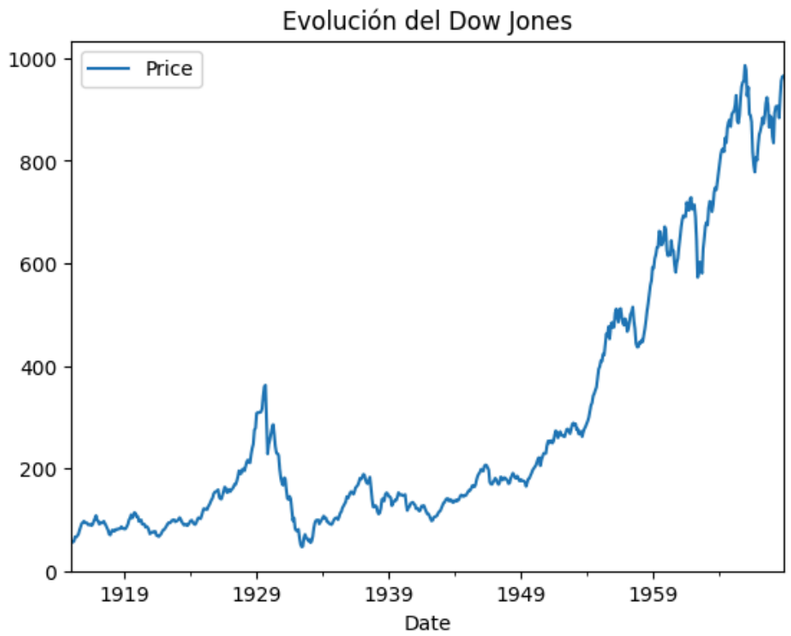
- Muestra la relación entre una variable independiente (Eje X) y una variable dependiente (Eje Y).
- Permite mostrar relaciones lineales o no-lineales.
- Correlaciones.
- Outliers.

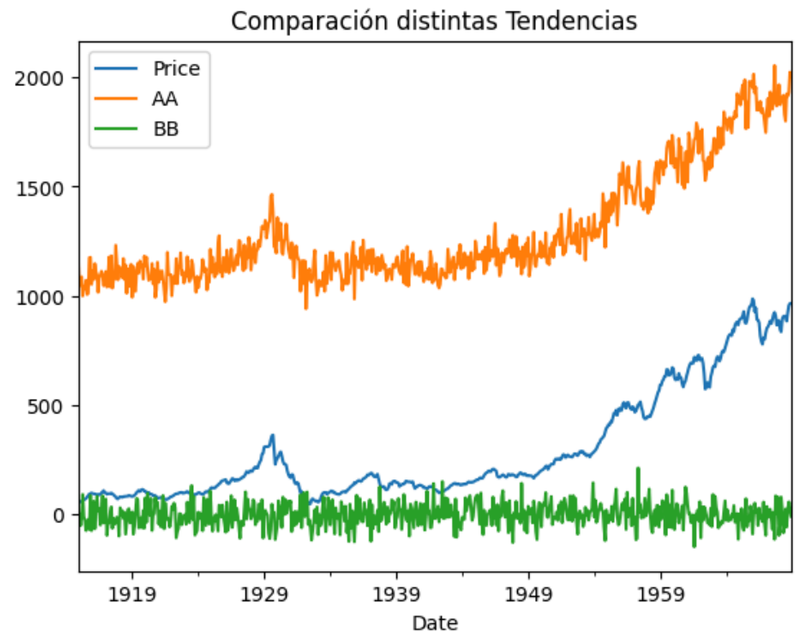


Visualizaciones: Líneas

Lineplot

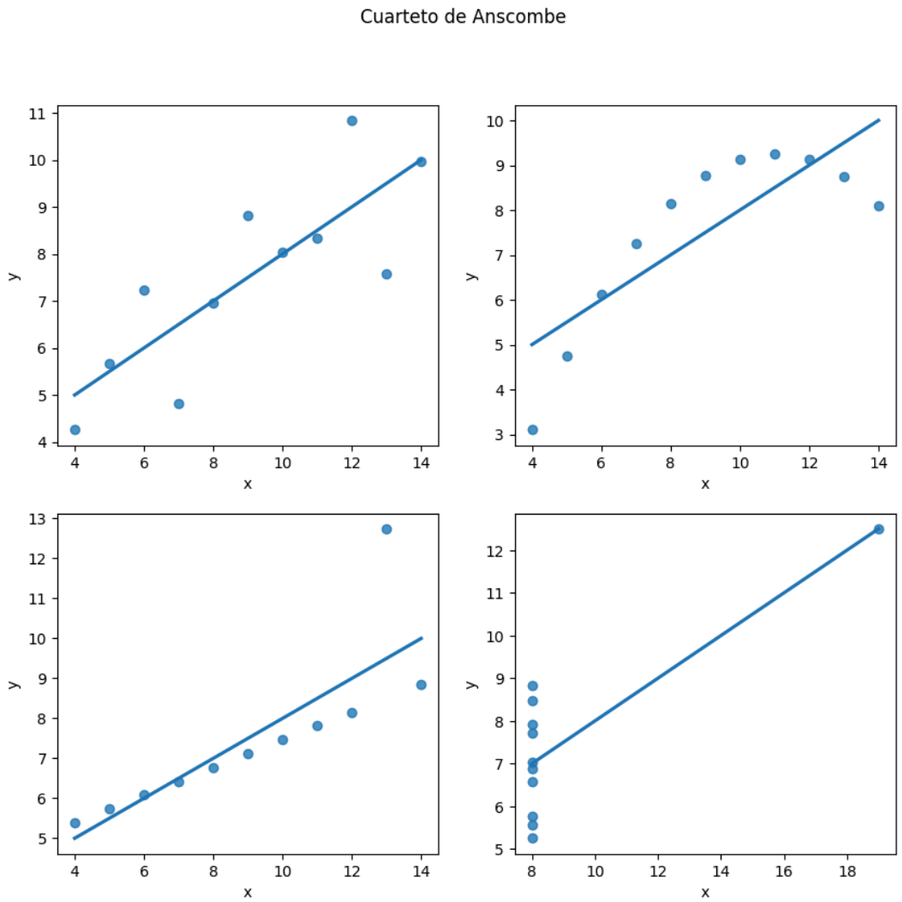
> Gráfico empleado para visualizar tendencias y su evolución en el tiempo.





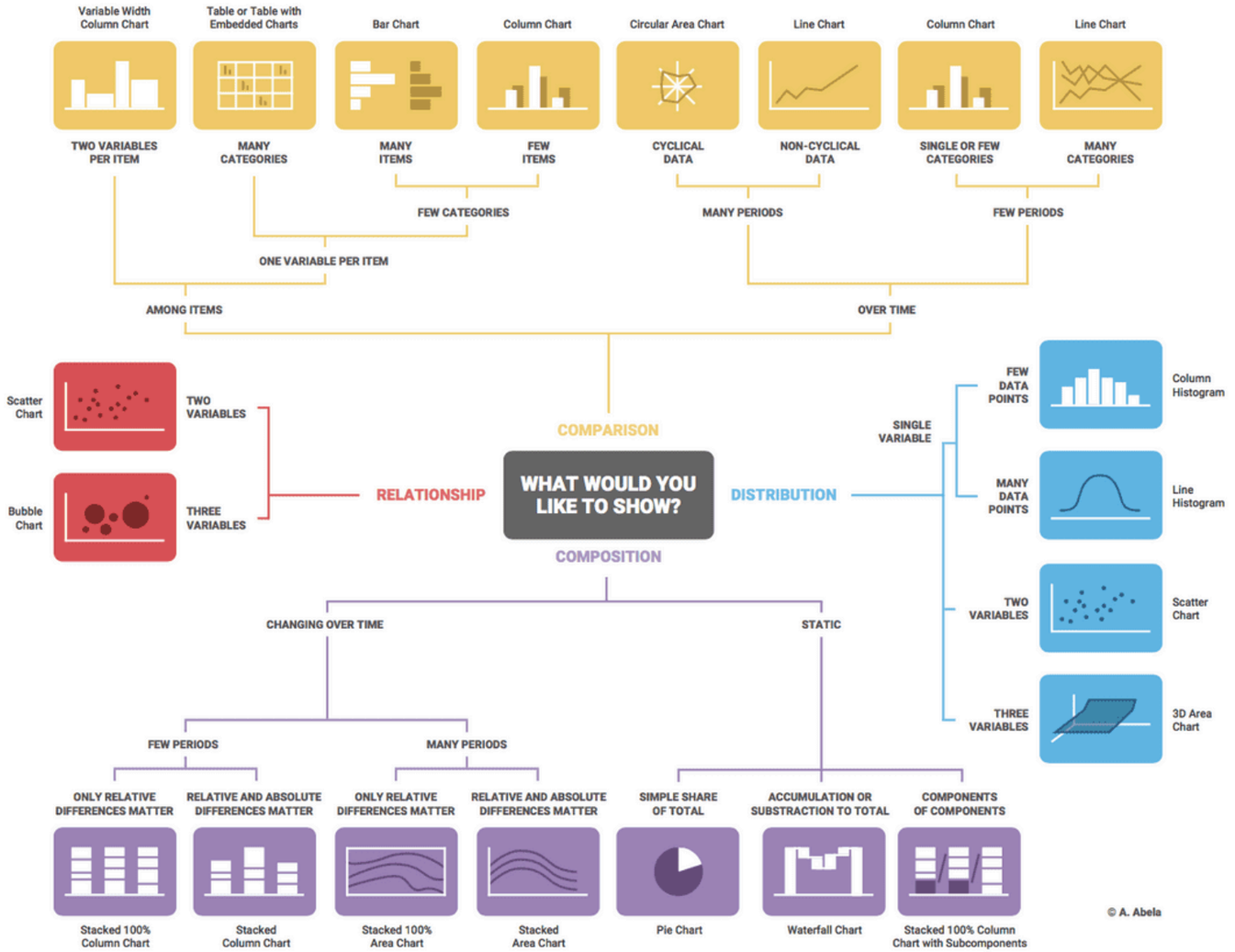
- Si bien es posible utilizarlo para gráficar dos medidas continuas, las buenas prácticas indican que el eje X siempre debería contener una componente temporal.

Estadísticos vs Visualizaciones



Propiedad	Valor
Media de cada una de las variables *x*	9.0
Varianza de cada una de las variables *x*	11.0
Media de cada una de las variables *y*	7.5
Varianza de cada una de las variables *y*	4.12
Correlación entre cada una de las variables *x* e *y*	0.816
Recta de regresión	$y = 3 + 0.5x$

¿Otras Visualizaciones?



Preguntas para terminar

- ¿Por qué usar visualizaciones? ¿Qué son los canales visuales?
- ¿Por qué es necesario el EDA?
- ¿Por qué es necesario utilizar tanto Estadísticos como Visualizaciones?

Le cours est terminé