



Minería de Datos

Análisis Exploratorio de Datos

Por
Elizabeth León Guzmán, Ph.D.
Profesora
Ingeniería de Sistemas
Grupo de Investigación MIDAS

Análisis Exploratorio



Análisis Exploratorio de Datos (AED)

Aproximación para el análisis de datos que emplea técnicas en su mayoría **gráficas** para:

- Entender un conjunto de datos.
- Descubrir estructuras subyacentes.
- Extraer variables importantes.
- Detectar valores atípicos y anomalías.
- Probar suposiciones subyacentes.



¡Una imagen vale más que mil palabras!

Técnicas cuantitativas

Son un conjunto de procedimientos estadísticos que proveen **salidas numéricas** o tabulares.

Ejemplos de técnicas cuantitativas incluyen:

- Pruebas de Hipótesis.
- Análisis de varianza.
- Puntos estimados e intervalos de confianza.
- Mínima regresión cuadrada.

¡Poco usadas en análisis exploratorio!

AED

Lo que distingue AED es el énfasis en las **técnicas gráficas** para ganar entendimiento lo que se opone a la aproximación clásica de las pruebas cuantitativas.

La mayoría de los analistas de datos usan una **mezcla** de técnicas gráficas y clásicas cuantitativas para encaminar sus problemas.

Técnicas de AED

- Graficar datos crudos como: histogramas, diagramas de dispersión.
- Graficar estadísticas simples como: gráficas de promedio, gráficas de desviación estándar, gráficas de cajas, etc.
- Organización de diagramas para maximizar la habilidad de reconocimiento natural de patrones (ej: utilización de múltiples gráficas por página)

Objetivos del AED

Maximizar el entendimiento del analista de un conjunto de datos, capturando la estructura subyacente. Items que un analista desea extraer del conjunto de datos:

- Lista de valores atípicos
- Interpretación robusta de conclusiones
- Estimados para parámetros
- Incertidumbres para esos estimados
- Lista ordenada de los factores importantes

Preguntas de Análisis



Identificar las preguntas relevantes para el problema.

- Se necesita identificar cuales preguntas necesitan ser respondidas y cuales preguntas no tienen que ver con el problema.
- Se debe dar prioridad a las preguntas en orden decreciente de importancia.
- Las técnicas de AED están vinculadas a cada una de éstas preguntas.

Preguntas que responde AED

1. ¿Qué es un valor típico?
2. ¿Cuál es la incertidumbre para un valor típico?
3. ¿Cuál es la distribución que mejor se ajusta a un conjunto de números?
4. ¿Qué es un percentil?
5. ¿Tiene el factor algún efecto?
6. ¿Cuáles son los factores más importantes?
7. ¿Cuál es la mejor función para relacionar una variable de respuesta a un conjunto de factores de variables?
8. ¿Cuáles son los valores óptimos para los factores?
9. ¿Se puede separar la señal del ruido en datos dependientes del tiempo?
10. ¿Podemos extraer alguna estructura desde datos multivariados?
11. ¿En los datos hay datos atípicos?

Estadísticas Descriptivas

Son números que resumen las propiedades de los datos :

Ubicación: promedio, mediana, moda.

Dispersión: desviación estándar.

La mayoría de las estadísticas descriptivas pueden ser calculadas en un solo paso a través de los datos.

Frecuencia y moda

La **frecuencia** de un valor de un atributo es el porcentaje de veces que el valor aparece en un conjunto de datos.

Por ejemplo, dado un atributo género y una población representativa de gente, el género femenino aparece alrededor del 50% de las veces.

La **moda** es el valor del atributo más frecuente

Las nociones de frecuencia y la moda son usadas típicamente con datos categóricos.

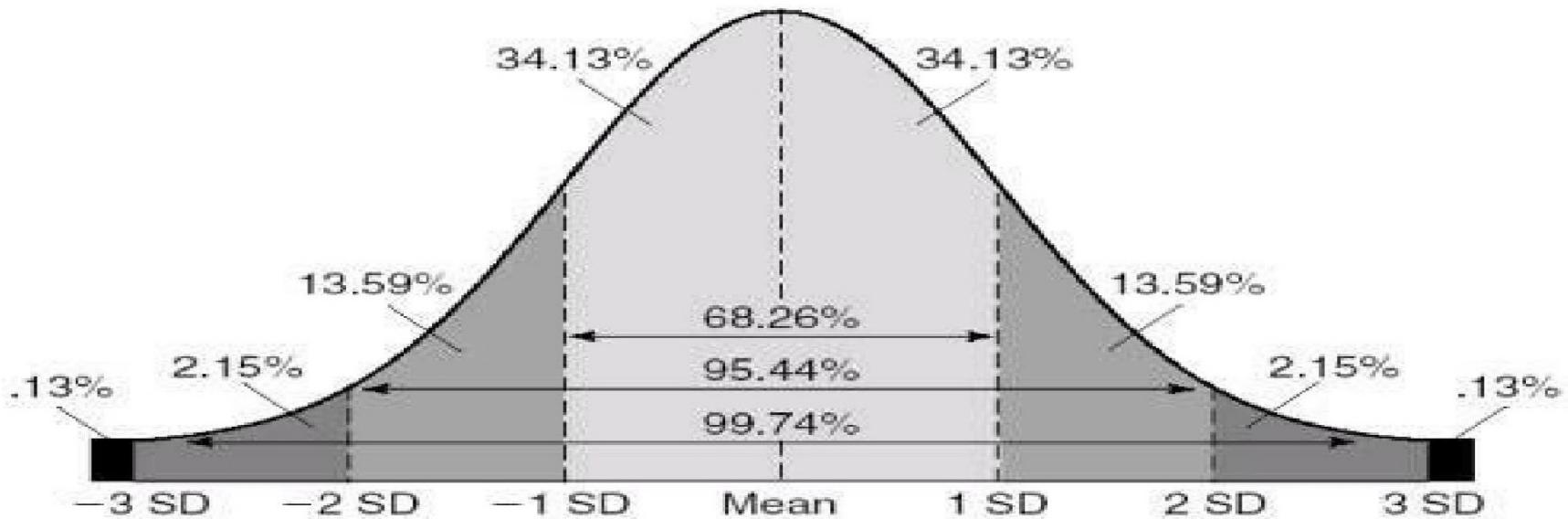
Percentiles

La noción de un percentil es más útil para datos **continuos**.

Dado un atributo continuo x y un número p entre 0 y 100, el p percentil es un valor x_p de x tal que el $p\%$ de los valores observados de x son menores a x_p

Para el 50 percentil es el valor de $x_{50\%}$ donde el 50% de todos los valores de x son menores a $x_{50\%}$

Percentiles



■ FIGURE 15.8 Percentile ranks and standard scores in relation to the normal curve.
SD = standard deviation.

Medidas de Ubicación Media y Mediana

□

La media es la medida más común de la ubicación de un conjunto de datos.

La media es muy sensible a los datos atípicos.

La mediana es comúnmente usada

$$\text{mean}(x) = \bar{x} = \frac{1}{m} \sum_{i=1}^m x_i$$

$$\text{median}(x) = \begin{cases} x_{(r+1)} & \text{if } m \text{ is odd, i.e., } m = 2r + 1 \\ \frac{1}{2}(x_{(r)} + x_{(r+1)}) & \text{if } m \text{ is even, i.e., } m = 2r \end{cases}$$

Medidas de dispersión: Rango y Varianza

Rango es la diferencia entre el máximo y el mínimo

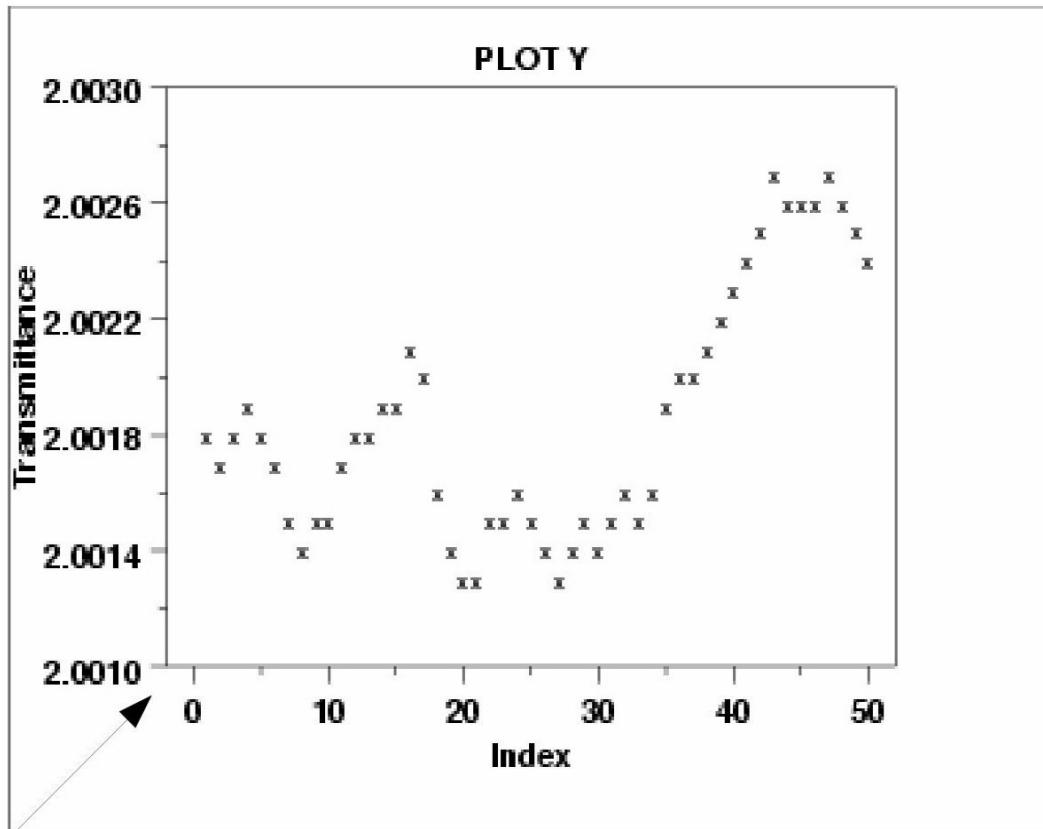
La varianza o desviación estándar es la medida más común de dispersión de un conjunto de puntos.

$$\text{variance}(x) = s_x^2 = \frac{1}{m-1} \sum_{i=1}^m (x_i - \bar{x})^2$$

Gráfica de Ejecución de Secuencia

Puede ser usada para responder las siguientes preguntas:

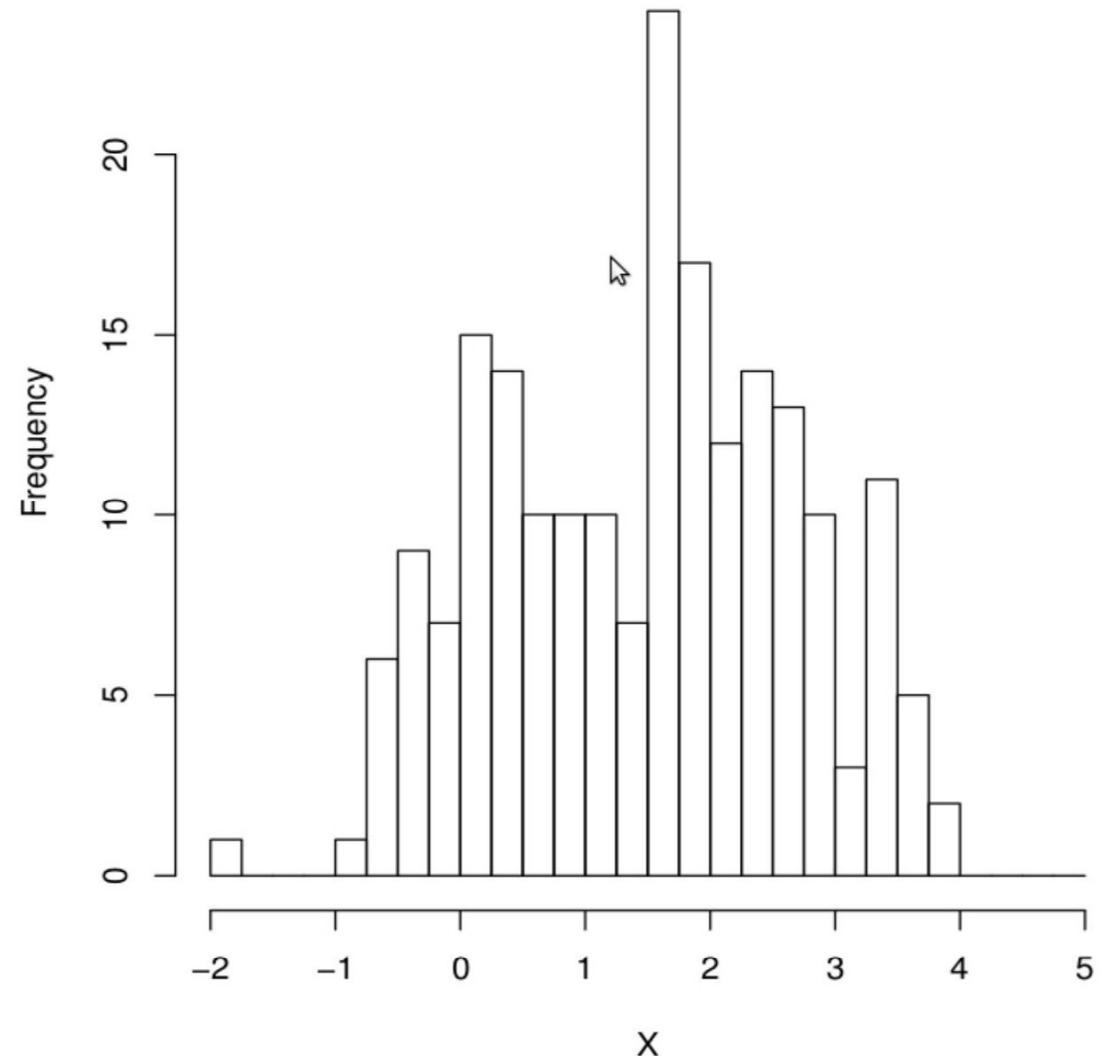
- ¿Hay cambios de localización?
- ¿Hay cambios en la variación?
- ¿Hay cambios en la escala?
- ¿Hay cambios en los valores atípicos?



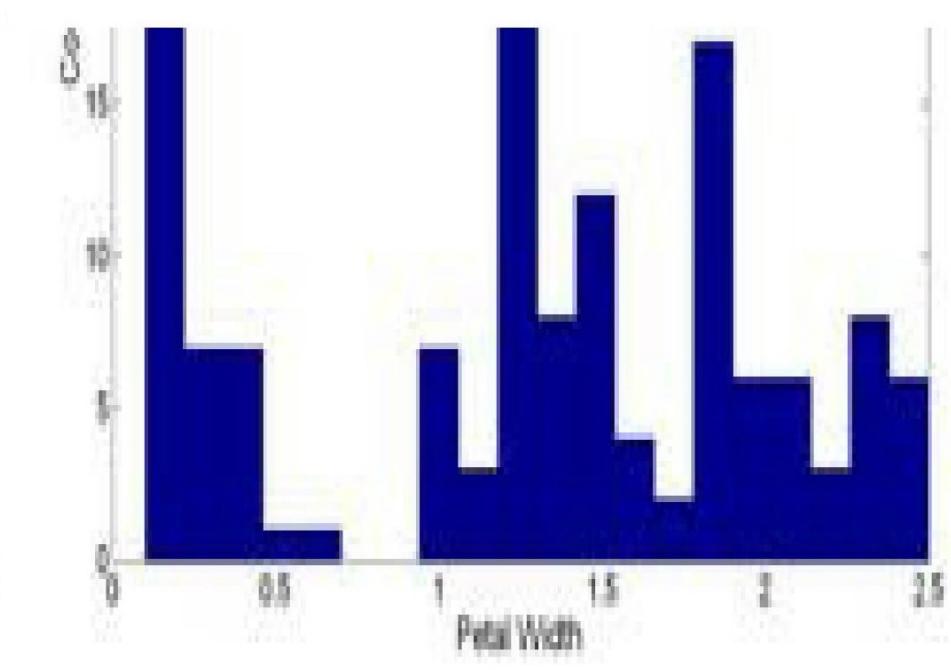
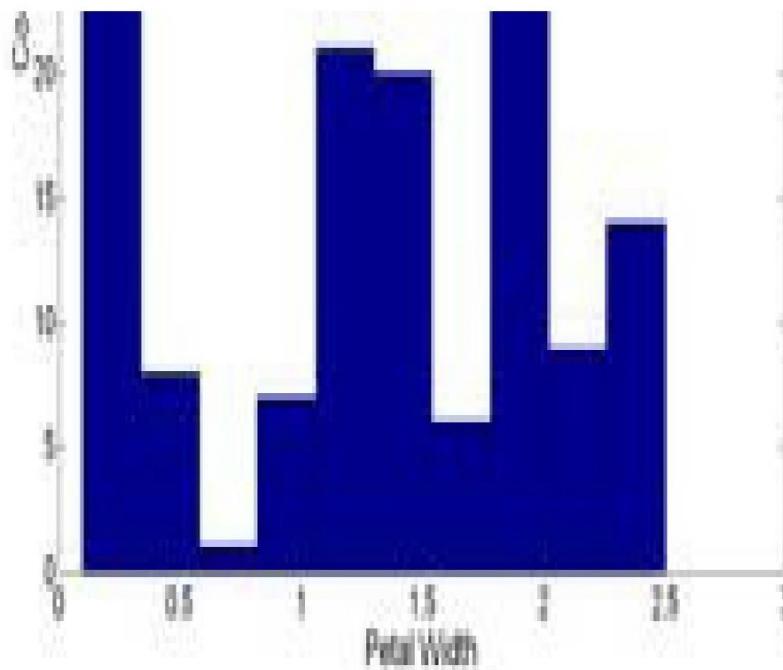
Eje y → Variable de respuesta

Histogramas

- Muestra la distribución usual de valores de una variable simple.
- Divide los valores en clases y muestra una gráfica de una barra del número de objetos en cada clase.
- La altura de cada barra indica el número de objetos.
- La forma del histograma depende del número de clases



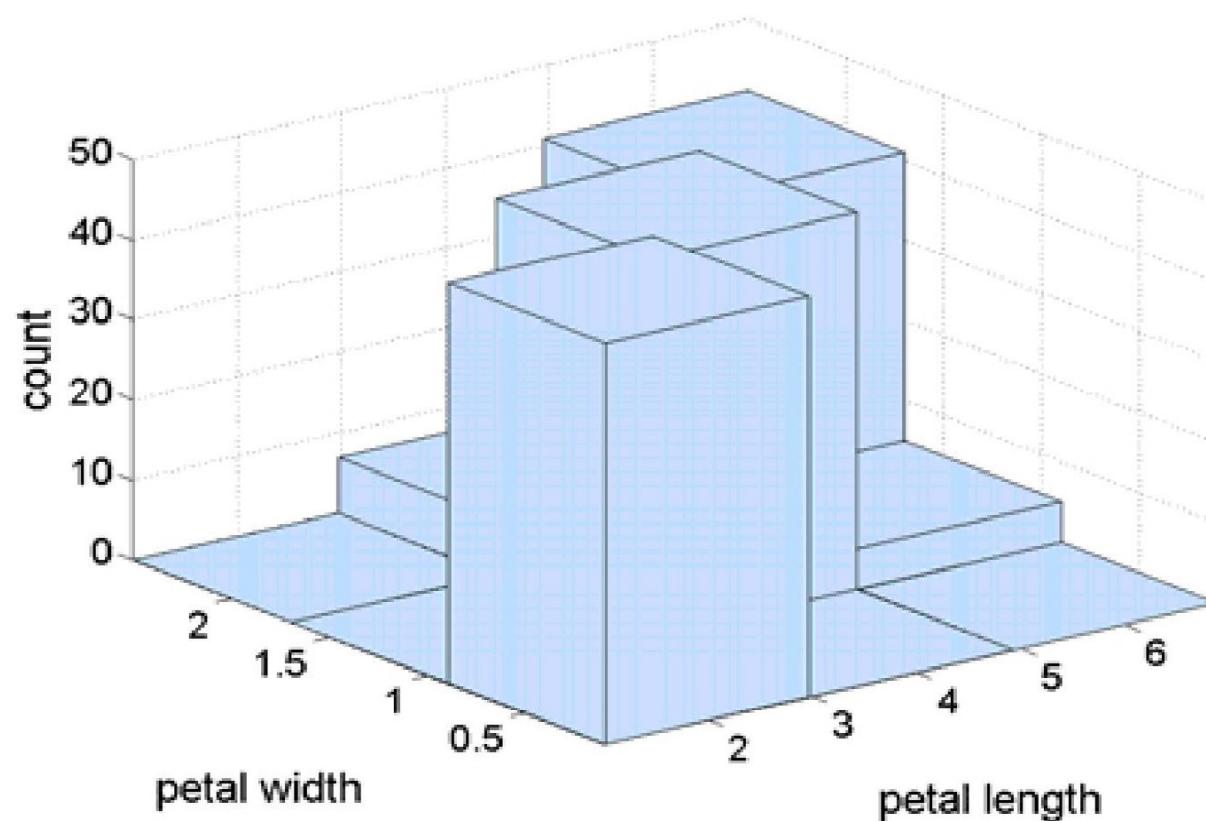
Ejemplo: Anchura del pétalo (10 y 20 clases, respectivamente)



Histogramas de dos dimensiones

Muestra la unión de distribución de valores de dos atributos.

Ejemplo: Anchura del pétalo y largo del pétalo ¿Qué nos está diciendo?



Histograma

- El histograma puede ser usado para responder las siguientes preguntas:
 - ¿De qué tipo es la distribución de la población de donde vienen los datos?
 - ¿Dónde están ubicados los datos?
 - ¿Son los datos simétricos o asimétricos?
 - ¿Hay valores atípicos?

Gráfica de Dispersión

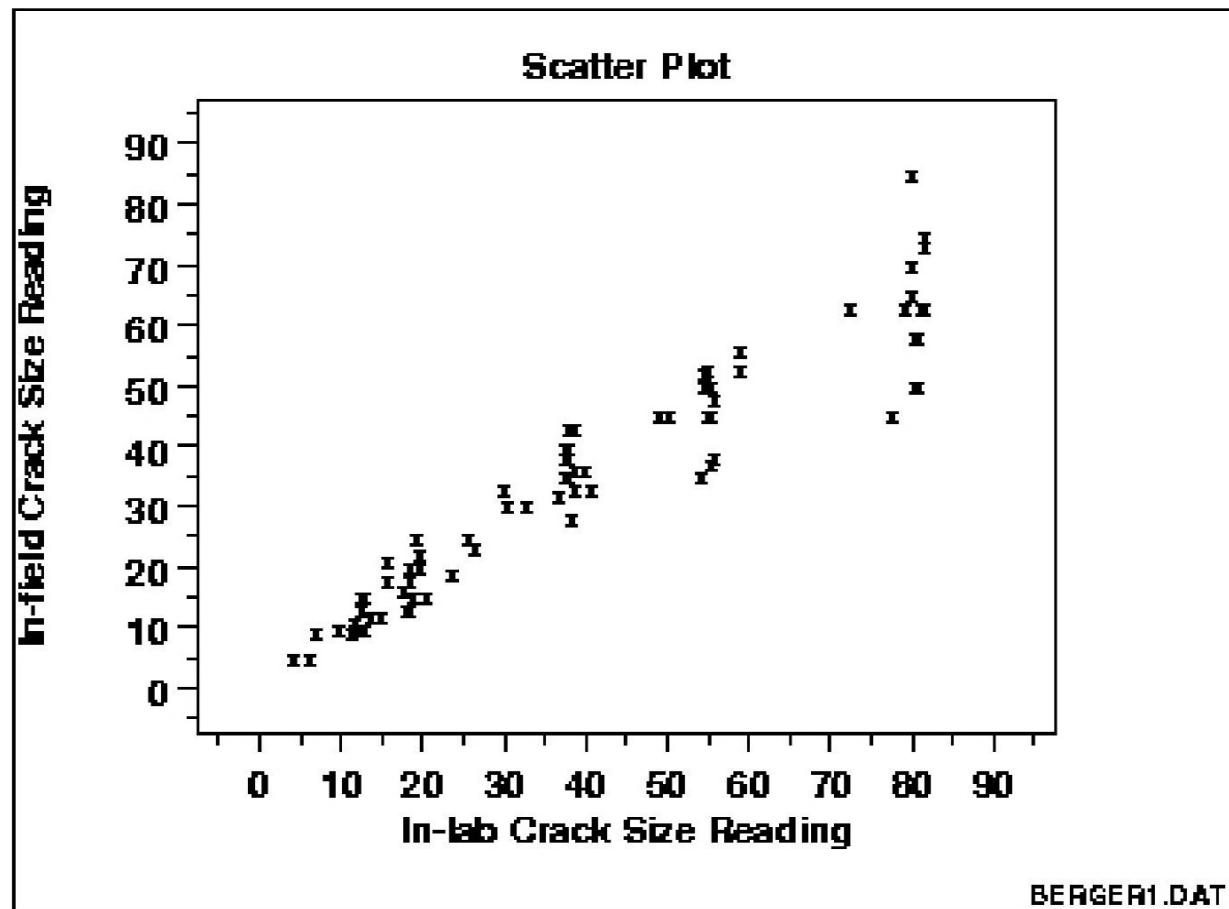
Propósito: verifica la relación entre variables.

Un gráfico de dispersión revela la relación o asociación entre dos variables.

Como las relaciones se manifiestan por si solas por una estructura no aleatoria en la gráfica.

Gráfica de Dispersión

Revela la relación lineal entre dos variables indicando que un *modelo de regresión* lineal puede ser apropiado.



Gráfica de Dispersión

Preguntas:

¿Están las variables X y Y relacionadas?

¿Están las variables X y Y relacionadas linealmente?

¿Están las variable X y Y relacionadas de manera no lineal?

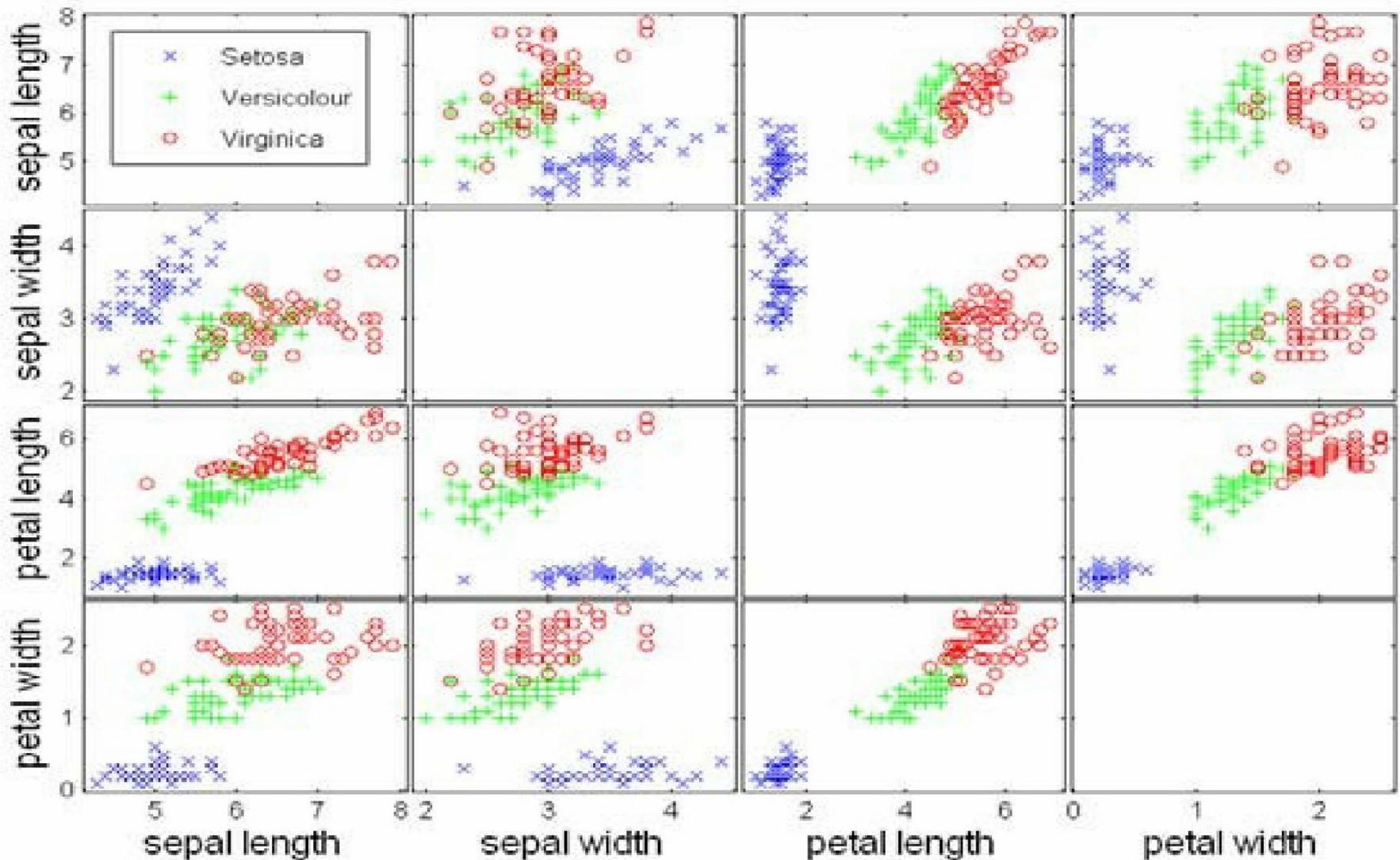
¿La variación de Y cambia dependiendo a X?

¿Hay valores atípicos?

Gráficas de Dispersion

Combinación en múltiples gráficas: Las gráficas de dispersión pueden ser combinadas en múltiples gráficas por página para ayudar a entender la estructura en un nivel más alto en conjuntos de datos con más de dos variables.

Scatter Plot Array of Iris Attributes



Gráfica de Dispersión

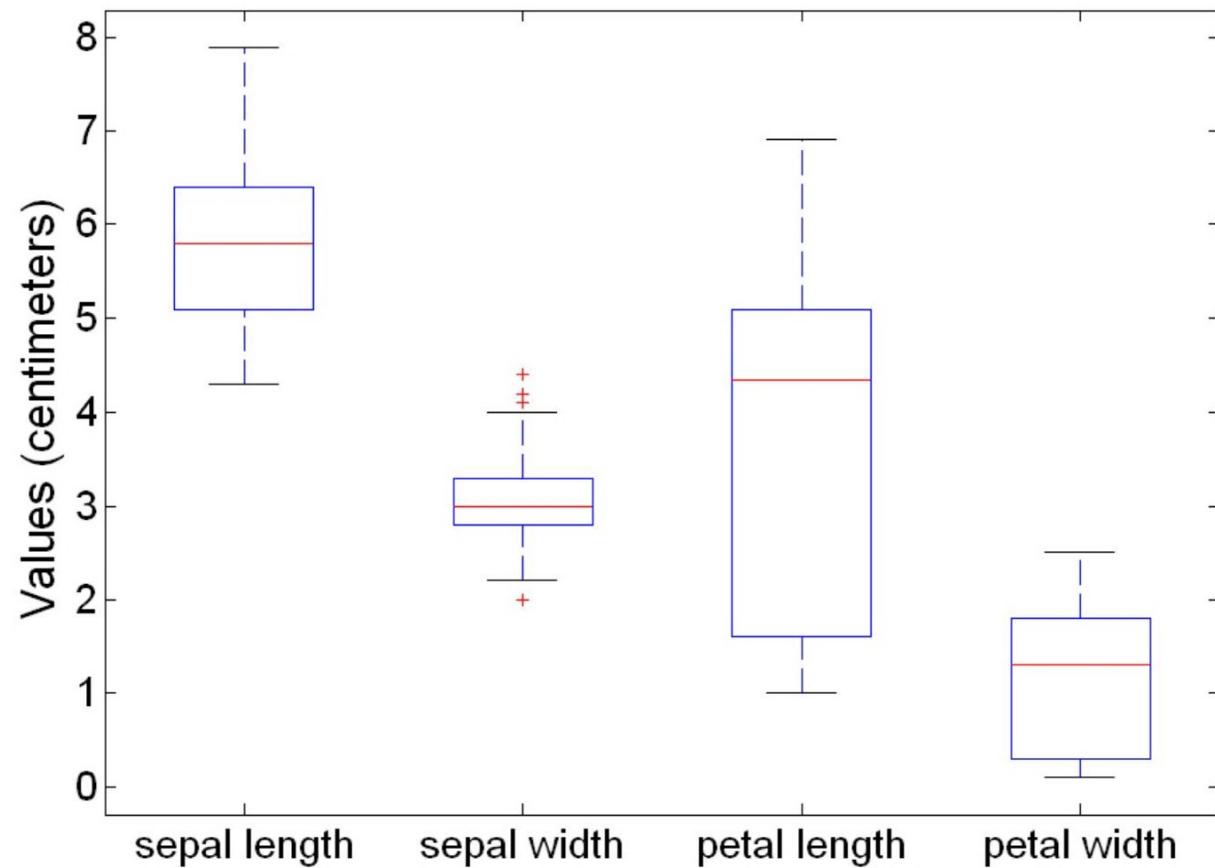
La gráfica de dispersión descubre las relaciones en los datos.

- Relaciones significa que hay algunas asociaciones estructuradas
 - (lineales, cuadráticas, etc.) entre X y Y.
 - **Se tiene en cuenta que:**
 - *Causalidad implica asociación
 - *Asociación no implica causalidad.

Gráficas de caja

Propósito: Verificar la ubicación y la variación de los cambios.

Herramienta para transmitir información de la ubicación y la variación: detecta e ilustra la ubicación y los cambios de variación entre diferentes grupos de datos.

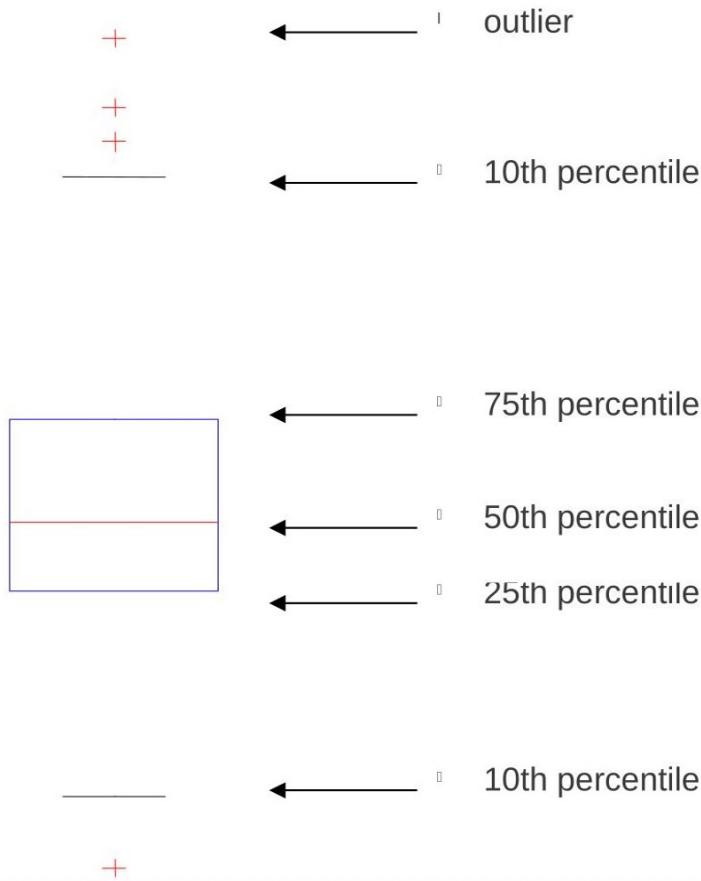


Gráficas de caja

□

Inventadas por J. Tukey.

Son otra forma de mostrar la distribución de los datos.



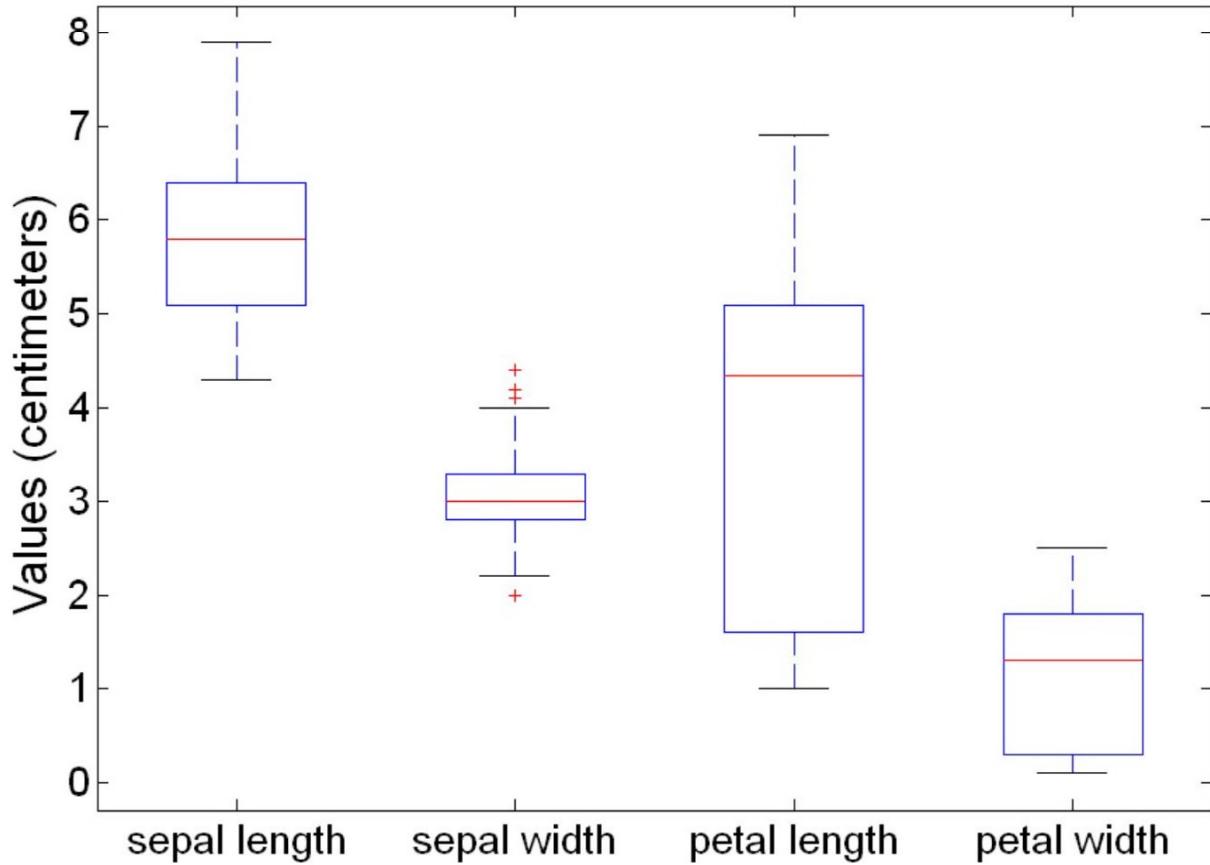
Gráficas de caja

Las gráficas de caja se conforman por:

Eje Vertical: valores del factor.

Eje Horizontal: El factor de interés

- Se calcula la mediana y los quartiles (el menor quartil es el percentil 25 y el quartil superior es el percentil 75)
- Se grafica un símbolo en la mediana (o una línea) y se dibuja la gráfica de caja entre el menor y mayor quartil; ésta caja representa el 50% de los datos.
- Se dibuja una línea desde el menor quartil al punto mínimo y otra línea desde el quartil mayor al máximo punto.



Literatura en AED

- Trabajo original “Exploratory Data Analysis”, **Tukey** (1977).
- Trabajos destacados:
 - Data Analysis and Regression, Mosteller and Tukey (1977),
 - Interactive Data Analysis, Hoaglin (1977),
 - El ABC's de EDA, Velleman y Hoaglin (1981).