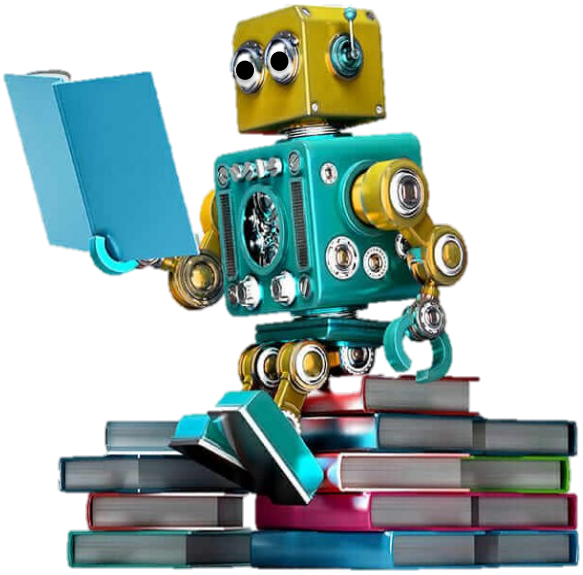
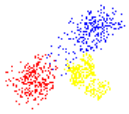
Preprocesamiento y Análisis de Datos

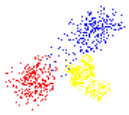




Data Preprocessing

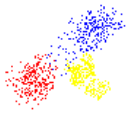
¿Por qué preprocesar los datos?

- **Datos incompletos:** falta de valores en algunos atributos, datos que vienen sólo agregados.
- **Datos ruidosos:** Errores de ingreso, outliers
- **Datos inconsistentes:** Diferencias en nombres de atributos para distintas áreas de la compañía, diferencias de unidades, codificaciones, mismo registro con distintos atributos en distintas bases de datos, etc.
- **Muchos datos:** A veces es necesario reducir la información para hacer el análisis



Análisis descriptivo de los datos

- **Objetivo:** Tener una visión de algunas características generales de los datos
- Es útil para el análisis inicial de la información.
- Ejemplo de algunas medidas descriptivas iniciales: media, mediana, varianza, etc.



Medidas de tendencia Central

- Media aritmética: $\overline{x} = \frac{\sum_{i=1}^{N} x_i}{N}$

- Media aritmética con pesos (weighted aritmetic mean):

$$\overline{x} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i}$$

- Algunos problemas: Sensibilidad a valores extremos

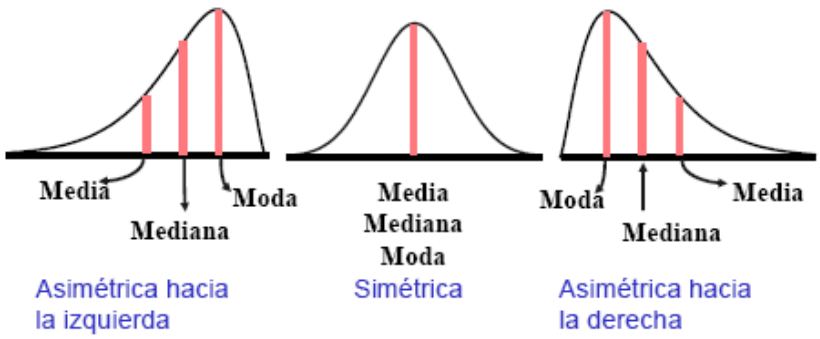
Medidas de tendencia Central (Cont..)

- **Media recortada (trimmed mean):** Se excluyen los valores más extremos para el cálculo de la media, ej. Se puede eliminar el 2% más alto y el 2% más bajo de los datos. Valores muy altos de exclusión causan pérdida de información

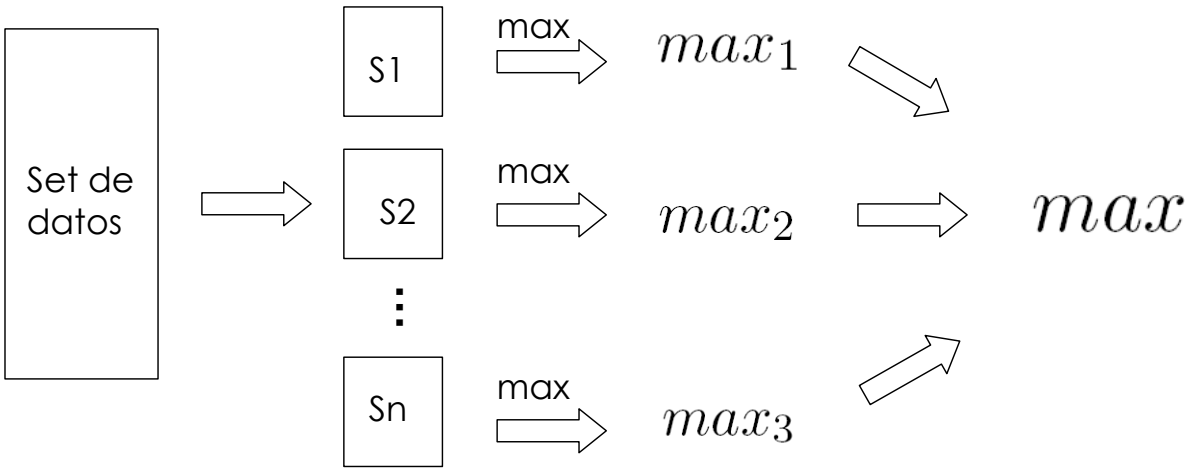
- **Mediana:** Utilizada más en datos asimétricos. Si ordenamos los números en un arreglo de tamaño N, la mediana corresponde al valor central del arreglo si N es impar, y al promedio de los dos valores del centro si N es par.

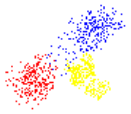
Medidas de tendencia Central (Cont..)

- Moda: Es el valor que tiene más frecuencia en el set de datos.



- Medidas distributivas: Medidas que se pueden obtener computando subconjuntos de los datos y luego mezclando los resultados (ej. sum,count,max,min)





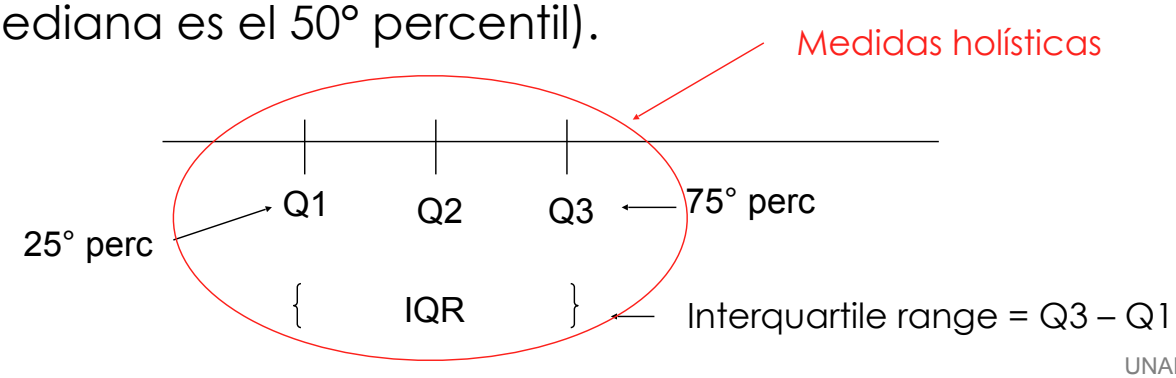
- **Medida holística:** Medidas que sólo se pueden obtener computando el set de datos completo como un todo. Ej Mediana.

Las medidas holísticas son más caras de computar

Medidas de Dispersión de los datos

- **Rango:** Para un set de datos $x_1, x_2, \ldots, x_N$ (Observaciones de un atributo), corresponde a la Diferencia entre el mayor y el menor valor

- **K-ésimo percentil:** Para un set de observaciones ordenadas en forma creciente corresponde al valor para el cual el K% de los datos queda antes que él (la mediana es el 50° percentil).

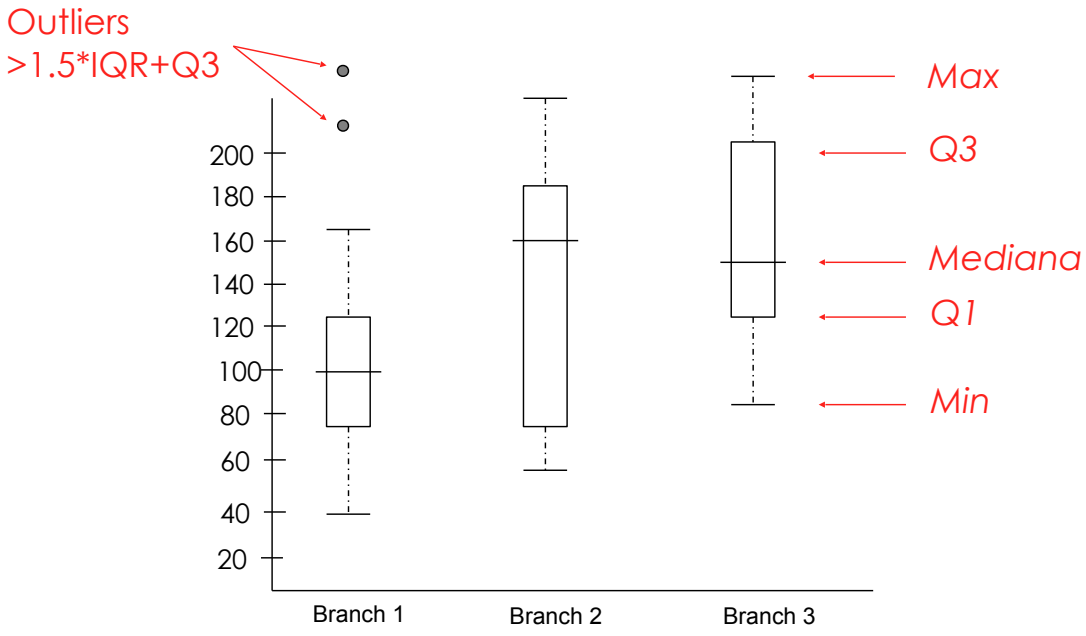


100 - quantiles = percentiles
10 - quantiles = deciles
5 – quantiles = quintiles
4 – quantiles = cuartiles

Ej. De uso de IQR: En detección de outliers, para valores a una distancia mayor a 1.5*IQR por sobre Q3 o bajo Q1.

- ***Five number summary:*** Corresponde a la secuencia de los valores *Min, Q1, Mediana, Q3, Max*

Boxplots: Forma gráfica de visualizar estos 5 valores:

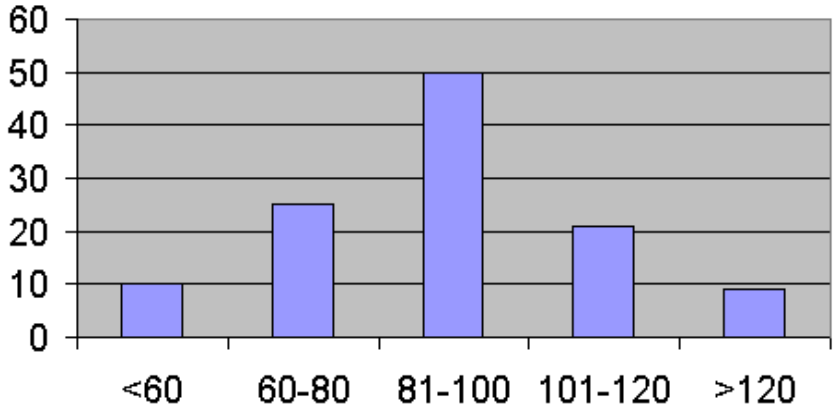


- **Varianza:** $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2$

σ = *desviación estándar*

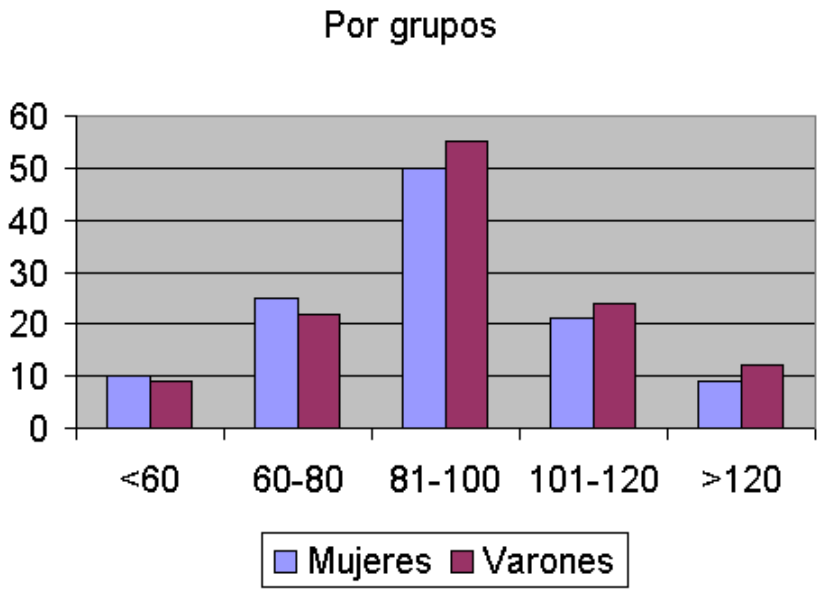
La desviación estándar es un indicador de la dispersión con respecto a la media cuando la media se está utilizando como medida central

Algunos Gráficos



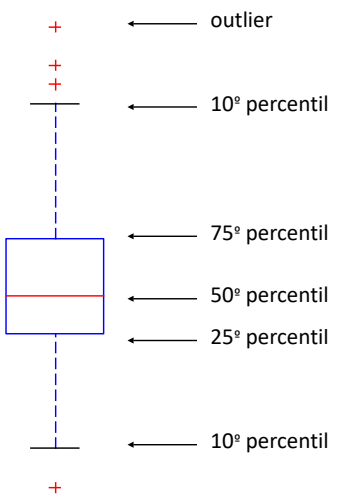
Histograma Simple

Algunos Gráficos(cont..)

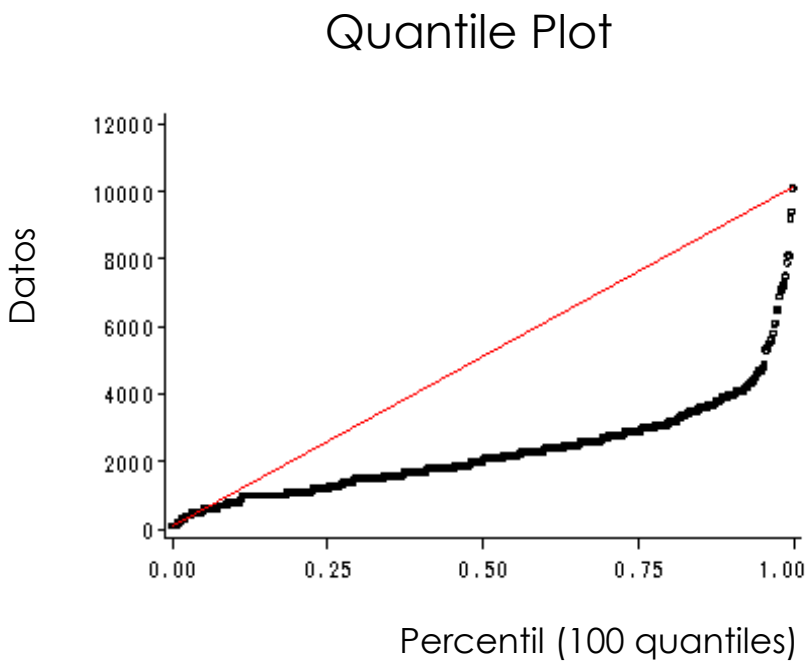


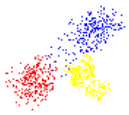
Técnicas de visualización: Box-plot

- Box-plot o diagramas de caja
  - Inventado por J. Tukey
  - Otra forma de mostrar la distribución de los datos
  - Siguiente figura muestra la parte básica de un diagrama de caja



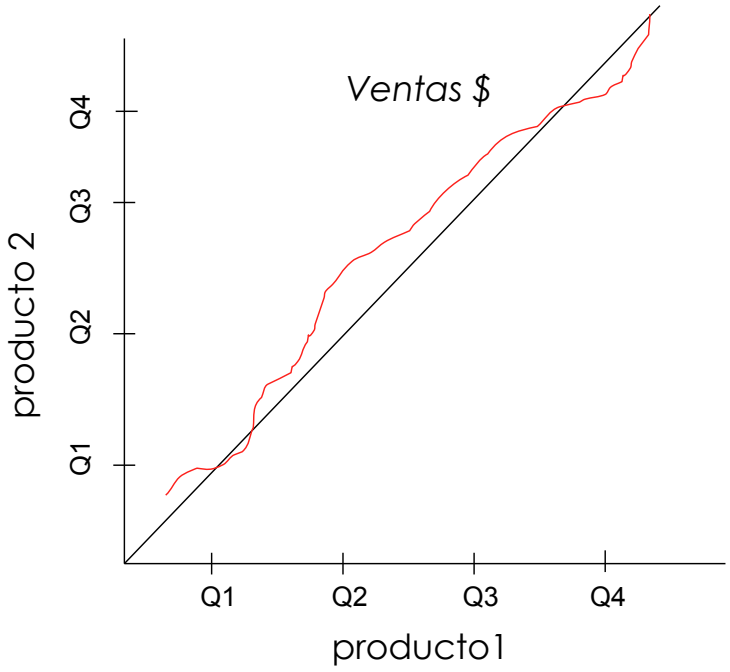
Algunos Gráficos(cont..)

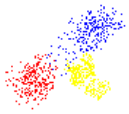




Algunos Gráficos(cont..)

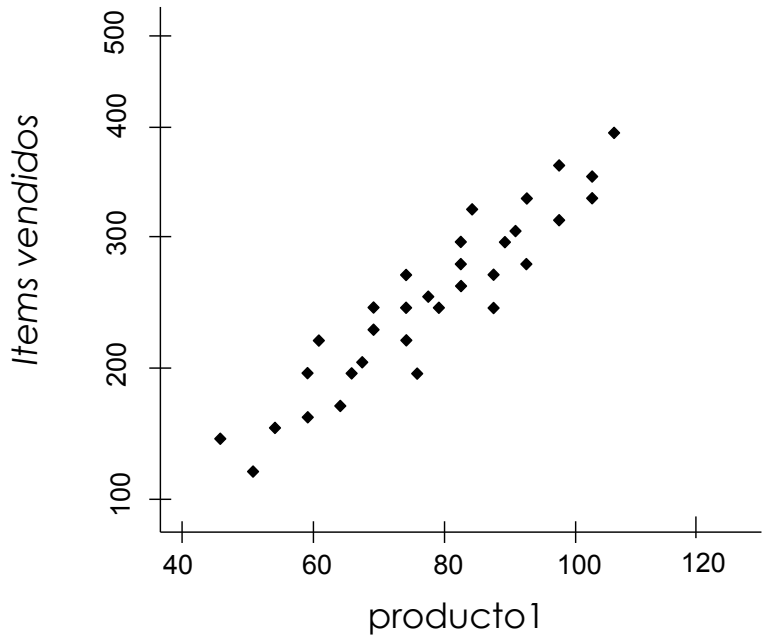
Quantile-Quantile Plot





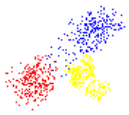
Algunos Gráficos(cont..)

Scatter Plot



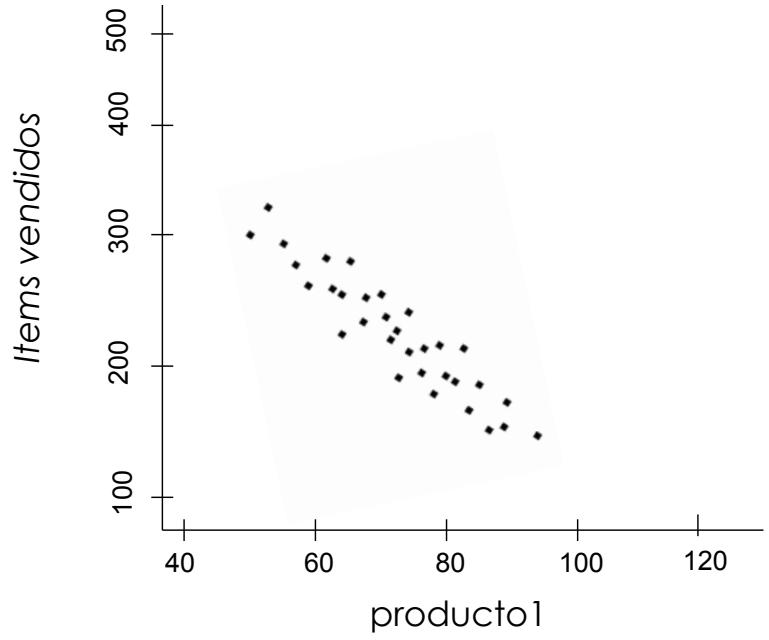
correlación positiva

Precio Unitario $

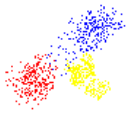


Algunos Gráficos(cont..)

Scatter Plot

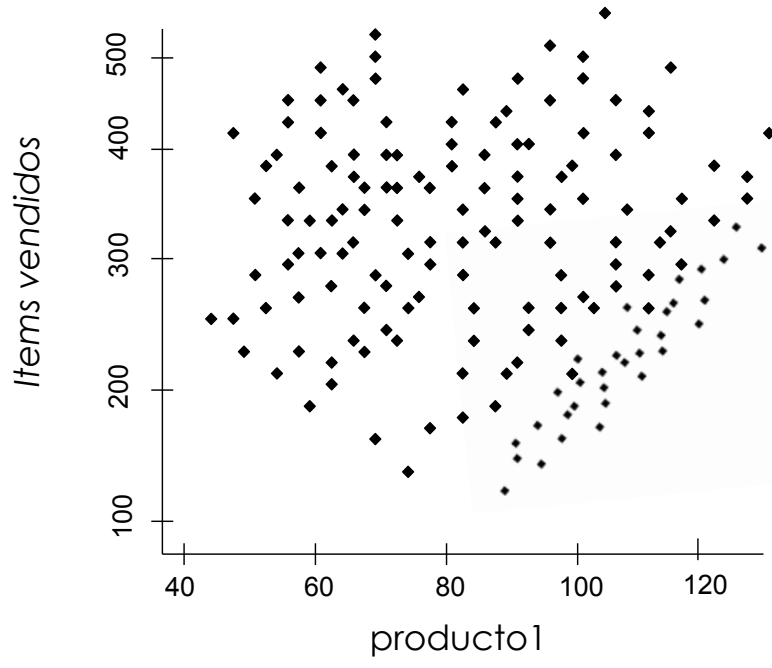


correlación negativa

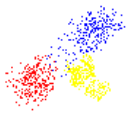


Algunos Gráficos(cont..)

Scatter Plot



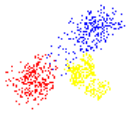
Sin correlación



Data Cleaning

- Datos en el mundo real tienden a ser incompletos, ruidosos e inconsistentes.

- El proceso de limpieza de datos trata de llenar los valores que faltan, identifica valores erróneos tratando de corregirlos y elimina inconsistencias en la información.

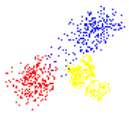


Datos Faltantes

- Muchas veces un atributo viene vacío

- Esta situación afecta el proceso de análisis

- Existen varias opciones para solucionar el problema:

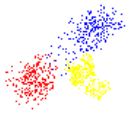


Datos Faltantes(Cont..)

- **Ignorar la tupla:** Método poco efectivo a menos que falten muchos atributos en la misma fila.
  Problemas cuando faltan valores en pocos atributos (aleatoriamente) pero en muchas tuplas .

- **Llenar los valores manualmente:** No es practicable cuando el set de datos presenta muchos valores faltantes

- **Usar una cte. Global para llenar los valores:** Ej: "desconocido", "-∞",etc. Trae problemas para algunos algoritmos de data mining que considerarían estos valores como datos válidos y trataría de encontrar patrones para ellos

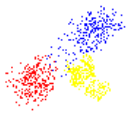


Datos Faltantes(Cont..)

- **Usar la media del atributo:** Llenar todos los valores faltantes en dicho atributo con el valor de la media para ese atributo. Es poco exacto

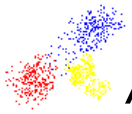
- **Usar la media por clases:** Igual que el método anterior pero utilizando la media considerando sólo los elementos que corresponden a la misma clase. Ej: Si falta el valor correspondiente al sueldo de un cliente de la clase business, llenarlo con el promedio del sueldo de todos los clientes business.

- **Usar el valor más probable:** Este valor puede ser determinado por regresión, herramientas de inferencia, árboles de decisión, etc.



Datos Faltantes(Cont..)

- No siempre un dato faltante es un error. Ej, persona no tiene licencia de conducir, no usa tarjeta de crédito, etc.
- En esos casos es importante tener valores definidos como “no se aplica”, etc.



Algunas técnicas de preprocesamiento

- **Binning:** Los datos se ordenan separándose en grupos (bins).
  - Smoothing by bin means: Cada valor en el bin es reemplazado por la media del bin.
  - Smoothing by bin boundaries: cada valor se reempla za por el valor mínimo del bin o el máximo dependiendo de cuál sea el más cercano.

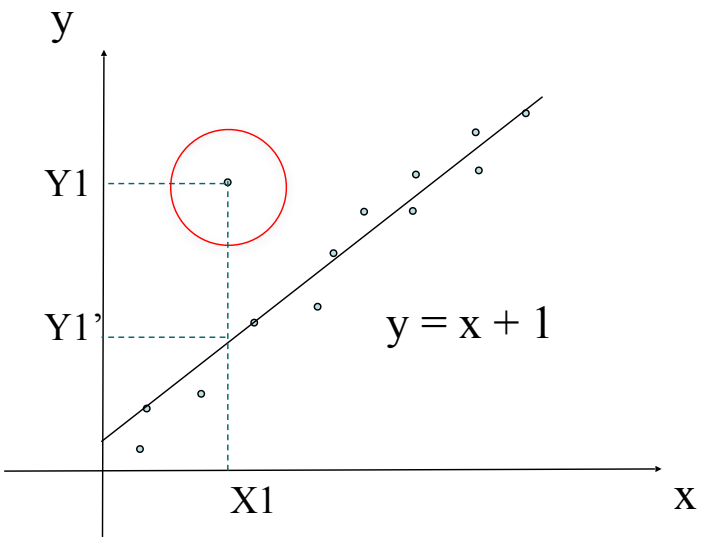
Este método se utiliza como herramienta de Discretización, smoothing, etc.

* Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
* Partition into (equi-depth) bins:
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34
* Smoothing by bin means:
  - Bin 1: 9, 9, 9, 9
  - Bin 2: 23, 23, 23, 23
  - Bin 3: 29, 29, 29, 29
* Smoothing by bin boundaries:
  - Bin 1: 4, 4, 4, 15
  - Bin 2: 21, 21, 25, 25
  - Bin 3: 26, 26, 26, 34

Algunas técnicas de preprocesamiento

•Regresión para corrección: Algunos datos se "corrigen" en base a una función. Ej:

Regresión Lineal



Algunas técnicas de preprocesamiento

- **Clustering** para la detección de outliers (candidatos a ser datos erróneos)