UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
FACULTAD DE CIENCIAS
ALMACENES Y MINERÍA DE DATOS

Pre-procesamiento de datos

Gerardo Avilés Rosas
gar@ciencias.unam.mx

Introducción

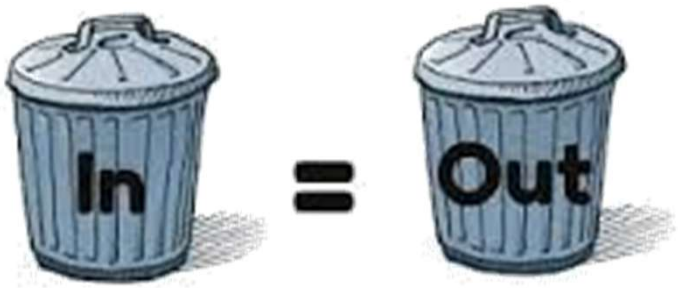
- Las **bases de datos** en teoría, no deberían albergar ***inconsistencias, datos sucios o valores faltantes***:
  - En grandes bases de datos, esto se complica, ya que en su mayoría, son el punto de convergencia de **fuentes de datos heterogéneas**.
  - Derivado de su gran tamaño y lo variado de los orígenes de datos, la **baja calidad de los datos aumenta**.
  - Las bases de datos no se diseñan para resolver problemas de análisis.

- Para resolver este tipo de problemáticas, es necesario realizar un **pre-procesamiento de datos**: tarea de vital importancia que debe ser llevada a cabo, tanto **BD** como en los **DWH**:
  - Aunque los **DWH** contienen **datos limpios e integrados**, esto no siempre significa que estén listos para entrar a un **algoritmo de minería de datos**.
  - Al igual que la **limpieza**, es un proceso que **consume mucho tiempo** y en la mayor parte de los casos no puede ser automatizado ya que la calidad es subjetiva.
  - Requiere la combinación de herramientas de cómputo y del analista.

…Introducción

- Como ya vimos, los datos en el mundo real están **sucios**:
  - **Incompletos:** carentes de valores en algunos atributos (o todos) o bien, carecen ciertos atributos de interés, o que contienen sólo datos agregados,
  - **Con ruido:** contiene errores (humanos o delibrados) o valores atípicos.
  - **Inconsistentes:** contiene discrepancias en códigos o nombres.

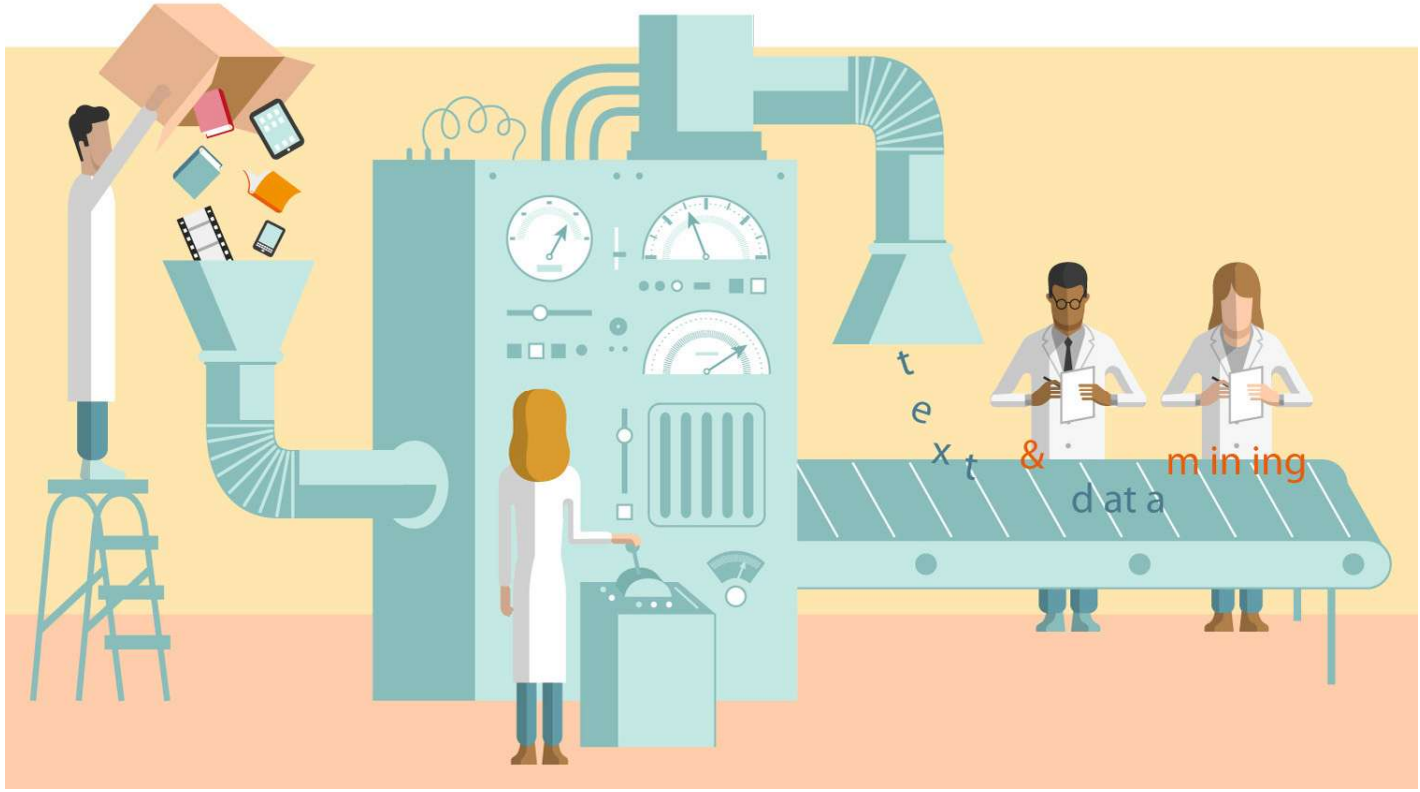
- De no llevar a cabo este proceso podemos obtener **resultados incorrectos** (pesimistas u optimistas):



  - **Si no hay datos de calidad → ¡No hay resultados de minado de calidad!**
  - **Decisiones de calidad** deben basarse en **datos de calidad**.
  - El pre-procesamiento es un proceso **no automatizado** que debe seguir un enfoque **iterativo incremental**.

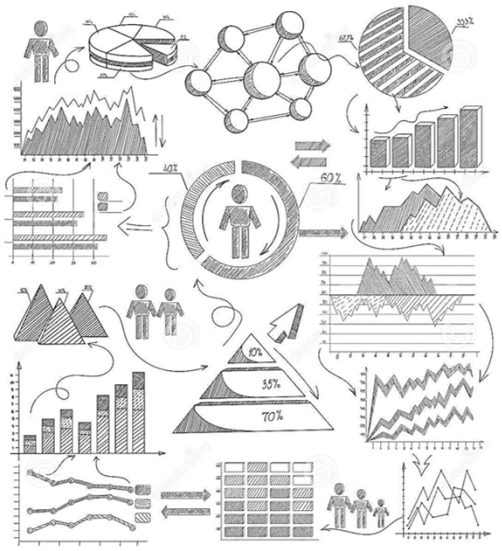
...Introducción

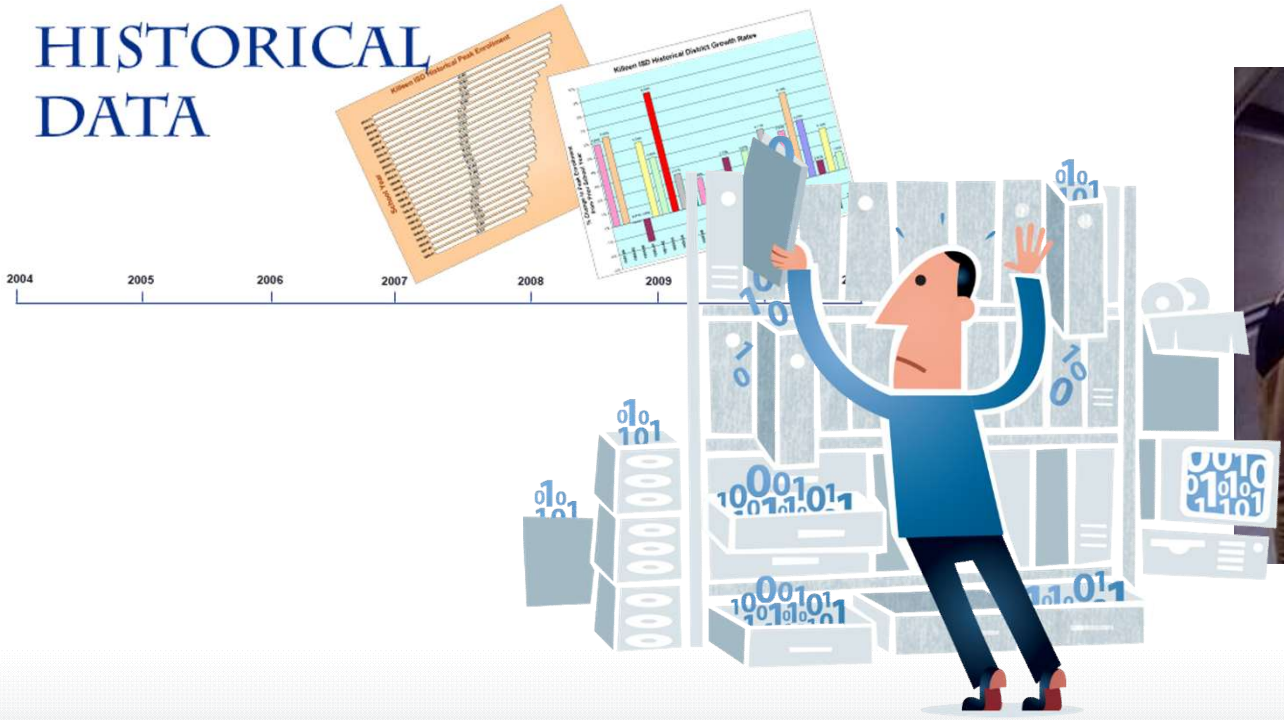
¿De qué forma pueden pre-procesarse los datos, de manera que podamos mejorar la calidad de los mismos?



Entendiendo los datos: Relevancia

- ¿Qué datos están disponibles para la tarea?
- ¿Todos los datos son relevantes?
- ¿Hay disponibilidad de datos relevantes adicionales?
- ¿Cuántos datos históricos están disponibles?
- ¿Quién es el experto de los datos?



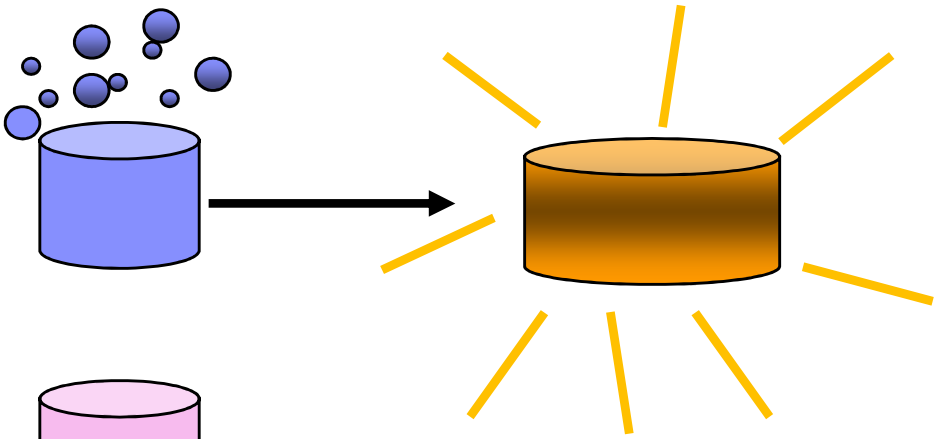


Entendiendo los datos: Cantidad

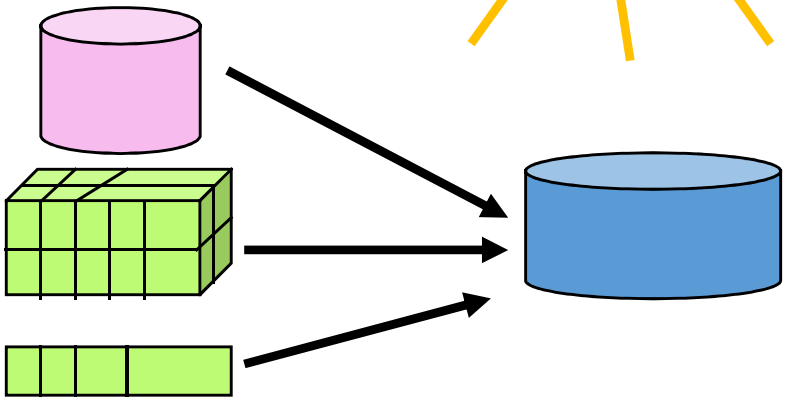
- Número de instancias (registros, objetos, etc.)
  - ***Regla general:*** *5,000 o más son deseables.*
  - *Si es menor, los resultados son menos fiables; utilizar métodos especiales (p.e.* ***boosting****).*
- Número de atributos
  - ***Regla general:*** *para cada instancia, alrededor de 10 atributos*
  - *Si hay más atributos, utilizar* ***reducción de características*** *y/o* ***selección****.*
- Número de instancias por clase
  - ***Regla general:*** *>100 instancias por cada clase*
  - *Si está* ***muy desequilibrado****, utilizar un* ***muestreo estratificado****.*

Preprocesamiento: Principales tareas

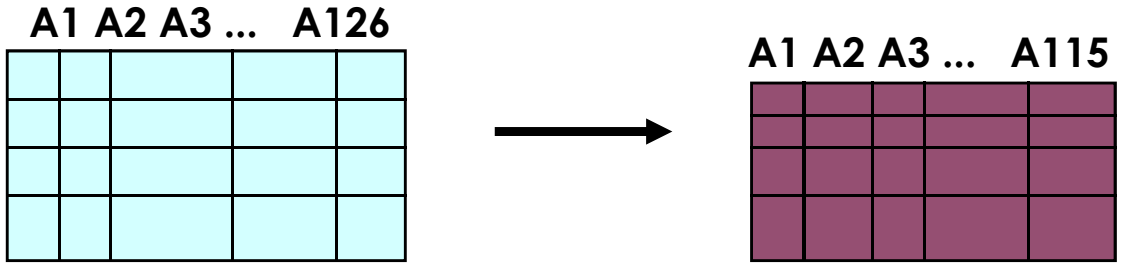
Limpieza de datos



Integración de datos



Reducción de datos



Transformación de datos -2, 32, 100, 59, 48 → **-0.02, 0.32, 1.00, 0.59, 0.48**

Preprocesamiento: Principales tareas

- **Limpieza de datos**

  *Se aplica con el objetivo de rellenar los **valores perdidos**, remover el **ruido**, resolver las **inconsistencias** de los datos, identificar o eliminar los **valores atípicos**.*

- **Integración de datos**

  *Permite **mezclar datos provenientes de múltiples fuentes** de datos heterogéneas en un repositorio coherente de datos (DWH).*

- **Reducción de datos**

  *Tiene como objetivo **reducir el tamaño de los datos** (agregando o eliminado características redundantes o agrupando). Se obtiene una representación reducida en volumen, pero que produce los mismos resultados analíticos (o similares).*

- **Transformación de datos**

  *Se aplican para **escalar** los datos y lograr que caigan en un **rango pequeño** (0.0 a 1.0). Se logra a través de **normalizaciones** o cálculo de **agregaciones**.*

- **Discretización de datos**

  Forma parte de la reducción de datos pero con especial importancia para datos numéricos.

… Preprocesamiento: Principales tareas

- En principio se utilizan para **mejorar la precisión y eficiencia** de los algoritmos de minado, principalmente aquellos que involucran **medidas de distancia**.
- Las técnicas mencionadas **no son mutuamente excluyentes** y están diseñadas para que trabajen juntas.

Limpieza de datos

- ***La actividad de convertir datos de origen en datos de destino, sin errores, sin duplicados, sin inconsistencias, discrepancias.***

- **Principales tareas** de la limpieza de datos:
  - ❑ Adquisición de datos y metadatos
  - ❑ Rellenar los valores faltantes
  - ❑ Unificar los formatos de fecha
  - ❑ Convertir valores nominales a numéricos
  - ❑ Identificar valores atípicos y suavizar los datos ruidosos
  - ❑ Corregir los datos inconsistentes

Limpieza de datos: Tareas

- **Adquisición de datos y metadatos**
  - *Los datos pueden estar en un **SMBD**: **OBDC** o **JDBC***
  - *En un **archivo plano**: columna fija, separados por coma, etc.*
  - *Verificar el **número de atributos** antes y después*
  - *Para los **metadatos**: tipos de atributos (nominales, categóricos, numéricos, etc.), rol de los atributos (entrada, salida, id, etc.) y descripciones de los mismos.*
- **Reformateo (convertir los datos a un formato estándar)**
  - *Llenado de valores perdidos*
  - *Unificación de formatos de fecha*
  - *Conversión de datos nominales a numéricos (habilitar comparaciones)*
  - *Agrupación de datos numéricos (binning).*
  - *Identificación de valores atípicos y suavizar los datos ruidosos.*
- **Corregir los datos inconsistentes**

Valores perdidos

- **Los datos no siempre está disponibles:**
  - *p. e. muchas tuplas no tienen valor registrado para varios atributos, como los ingresos de clientes en los datos de ventas.*
- **Los datos faltantes puede deberse a:**
  - *Mal funcionamiento de los equipos que los recolectan.*
  - *Inconsistencias con otros datos registrados y por lo tanto eliminados.*
  - *Los datos no se ingresan debido a que no se entienden.*
  - *Algunos datos pueden no ser considerados importantes en el momento de la entrada.*
  - *No se registra la historia o cambio de los datos.*
  - *Restricciones default para datos recolectados*
  - *Consideraciones diferentes entre el momento en que fueron recopilados y cuando fueron analizados.*
- **Los datos faltantes pueden necesitar ser inferidos**

…Valores perdidos

1. **Ignorar la tupla**

   Es un método que suele utilizarse cuando el valor del atributo no se tiene disponible. Es un método efectivo a menos que la tupla contenga varios atributos con valores perdidos y es extremadamente pobre cuando el porcentaje de valores perdidos varia en relación a los atributos por cada tupla.

2. **Llenar el valor manualmente**

   Este enfoque consume grandes cantidades de tiempo y no es factible cuando se tienen grandes bases de datos con varios valores perdidos.

3. **Usar una constante global**

   Se reemplazan todos valores perdidos de los atributos con el mismo valor constante, (p.e. **"Desconocido"**), sin embargo, corremos el riesgo de que los algoritmos de minería de datos erróneamente lo consideren un concepto interesante. Técnica simple pero no infalible.

4. **Utilizar una medida de tendencia central (imputación)**

   Se puede utilizar una medida que indique el valor "medio" de la distribución de datos. Para una **distribución normal** podemos utilizar la **media**, pero si la distribución esta **sesgada** se debería entonces usar la **mediana**.

…Valores perdidos

5. **Utilizar la media (o mediana) para todas las muestras que pertenezcan a la misma clase (imputación)**

   Por ejemplo, si deseamos hacer una clasificación de clientes por riesgo de crédito, es posible reemplazar los valores perdidos por la media (o mediana) del ingreso en el mismo riesgo de crédito.

6. **Utilizar el valor más probable**

   Este valor puede ser determinado con un **análisis de regresión**, o el **Teorema de Bayes**. Es posible también utilizar un **árbol de decisión**. En el ejemplo de la compañía se podría utilizar un árbol para predecir el valor perdido para ingreso.

…Valores perdidos

- Los **métodos 3 a 6** producen **valores sesgados** en los datos por lo que el valor no siempre podría ser correcto.
- El **método 6** es por el contrario una estrategia muy popular, pues en comparación con los otros métodos, **utiliza la mayor cantidad de información** del los datos actuales para predecir el valor perdido. Lo anterior nos brinda grandes oportunidades de encontrar la relación entre el atributo que andamos buscando y los demás valores presentes.
- Es importante tener presente que en algunos casos **los valores perdidos no implican necesariamente un error**.

Unificar los formatos de fecha

- Algunos sistemas aceptan fechas en diversos formatos:
  - Sept 24, 2016, 9/24/2016,24-09-2016, 24.09,2016, etc.
  - Las fechas internamente se transforman a valores estándar
- En la mayoría de las ocasiones, con el año es suficiente.
  - Es posible que se requiera más detalle: mes, día, trimestre, hora, etc.
- Se pueden utilizar formatos de representación alternativos:
  - **YYYYMM**, **YYYYMMDD**, etc., pero no preservan intervalos
  - Se pueden utilizar formatos como el de **UNIX** (número de segundos desde 1970) o el de **SAS** (número de días desde el 01/01/1960)
  - Utilizar formatos estandarizados:

$$KSP = YYYY + \frac{\text{días transcurridos desde el } 1^{\circ} \text{ de enero} - 0.5}{365\left(+1 \text{ si el año es bisciesto}\right)}$$

Conversión: nominal a numéricos

- Algunas herramientas pueden tratar con los valores nominales internamente:
  - Sin embargo, métodos como redes neuronales, métodos de regresión, vecino más cercano requieren trabajar solo con entradas numéricas.
  - En estos casos es necesario convertirlos a valores numéricos.
- Existen diferentes estrategias que abarcan: valores binarios, ordenados y atributos nominales multivaluados.
- **Valores binarios a numéricos**

  Género = [M, F]

  Se puede convertir a un campo de la forma **Campo_0_1** con valores **1 y 0:**

  P.e. Género = M → **Género_0_1 = 0**

  Género = F → **Género_0_1 = 1**
- **Atributos ordenados a numéricos** (p.e. las calificaciones) pueden convertirse a números, preservando el **orden natural**:

  A → 10.0 ; A- → 9.0 ; B+ → 8.0 ; B → 7.0

…Conversión: nominal a numéricos

- **Nominales (pocos valores)**, atributos multivaluados, no ordenados, con un número pequeño de valores (no más de 20)

  P.e. Color = rojo, naranda, amarillo,…,violeta

  Para cada valor *v se pueden crear un marcador binario en donde, si tiene el valor 1, es un color en particular, 0 en caso contrario*

ID	Color
371	Rojo
433	Amarillo

ID	C_rojo	C_naranja	C_amarillo	…
371	1	0	0	
433	0	0	1	

- **Nominal (muchos valores)**, por ejemplo, códigos para los estados o para profesiones:
  - ❑ Ignorar los campos ID ya que eso valores son únicos para cada registro.
  - ❑ Siempre que sea posible, formar grupos (regiones, seleccionar los más frecuentes)
  - ❑ Crear marcadores binarios para los valores seleccionados

…Conversión: nominal a numéricos

- **Nominal (muchos valores)**

Región	Estado
Noroeste	Baja California
	Baja California Sur
	Chihuahua
	Durango
	Sinaloa
	Sonora
Noreste	Coahuila
	Nuevo León
	Tamaulipas
Occidente	Colima
	Jalisco
	Michoacán
	Nayarit
Oriente	Hidalgo
	Puebla
	Tlaxcala
	Veracruz

Región	Estado
Centronorte	Aguascalientes
	Guanajuato
	Querétaro
	San Luis Potosí
	Zacatecas
Centrosur	Ciudad de México
	Estado de México
	Morelos
Suroeste	Chiapas
	Guerrero
	Oaxaca
Sureste	Campeche
	Quintana Roo
	Tabasco
	Yucatán

Datos con ruido

- ¿Qué es el ruido?

  El ruido es un error aleatorio o variación en una medida.

- Los **valores incorrectos (con ruido)** se pueden deber a:
  - *instrumentos de recolección de datos erróneos.*
  - *problemas de entrada de datos.*
  - *problemas de transmisión de datos.*
  - *limitación de la tecnología.*
  - *inconsistencia en convención de nomenclatura.*
- Otros problemas que requiere la limpieza de datos
  - *registros duplicados.*
  - *datos incompletos.*
  - *datos inconsistentes.*
- En los datos, se suelen emplear técnicas de **estadística descriptiva** o **métodos de visualización** para identificar dichos valores.

Discretización simple: *Binning*

- **Binning**

  Se trata de un método que atenúa el ruido en un valor ordenado, consultando su vecindario (valores alrededor de él). Los valores ordenados son distribuidos en ***bins****. Dado que este método consulta en el vecindario de valores, se lleva a cabo un suavizado local.*

- **Igual ancho** (distancia):
  - Se divide el rango en N intervalos de igual tamaño.
  - Sí A y B son el menor y el mayor valor del atributo, el ancho del intervalo se obtiene a partir de: $W = (B - A)/N$
  - El más directo, pero los valores atípicos pueden dominar la presentación
  - Los datos asimétricos no se manejan bien.

Datos ordenados: **4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34**

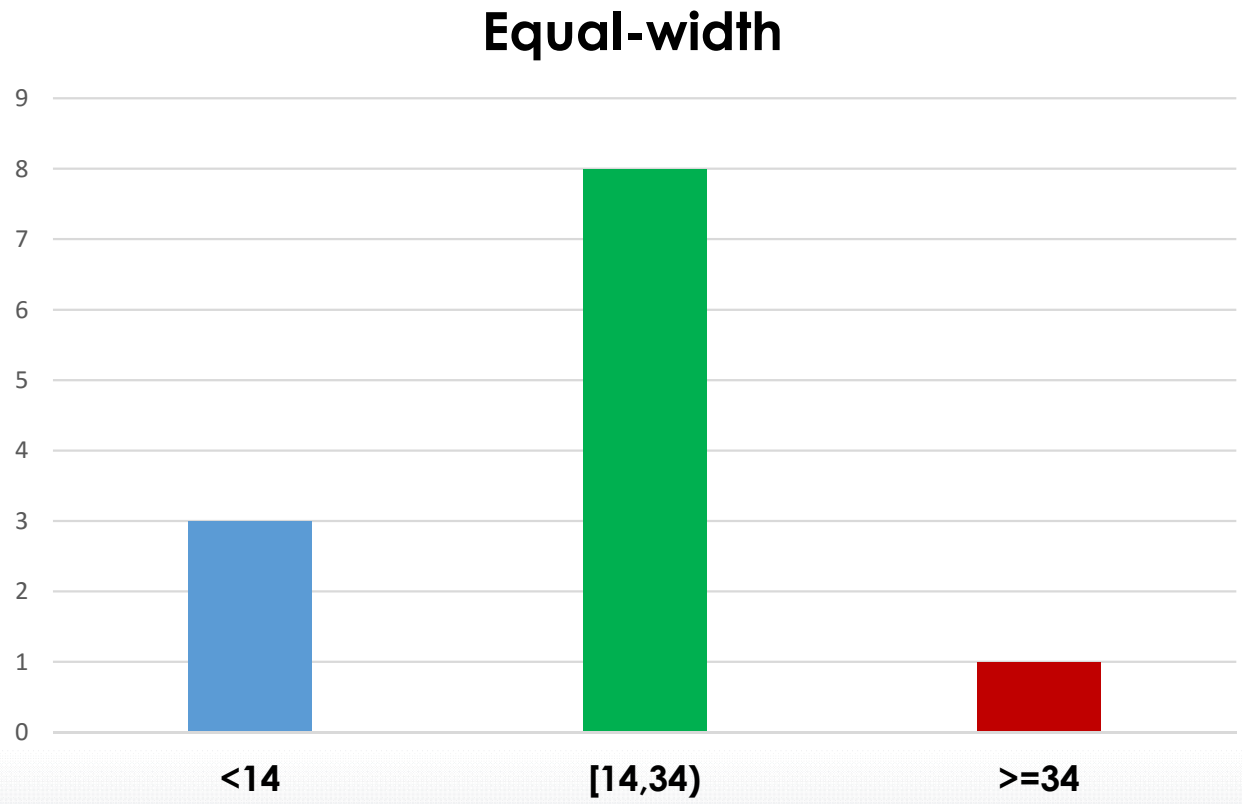
$$w = \frac{34 - 4}{3} = 10$$

Los límites del intervalo son:

$$\begin{cases} Bin_1 = \min + w = 4 + 10 = 14 \\ Bin_2 = \min + 2w = 4 + 20 = 24 \\ Bin_3 = \min + 3w = 4 + 30 = 34 \end{cases}$$

…Discretización simple: *Binning*

Bins de igual ancho		
Bin 1:	< 14	4, 8, 9
Bin 2:	[14,34)	15, 21, 21, 24, 25, 26, 28, 29
Bin 3:	>= 34	34



…Discretización simple: *Binning*

- **Igual profundidad** (frecuencia):
  - Se divide el rango en N intervalos, cada uno conteniendo aproximadamente el mismo número de muestras.
  - Buen escalamiento de datos
  - Administrar atributos categóricos puede ser complicado.

Datos ordenados: **4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34**

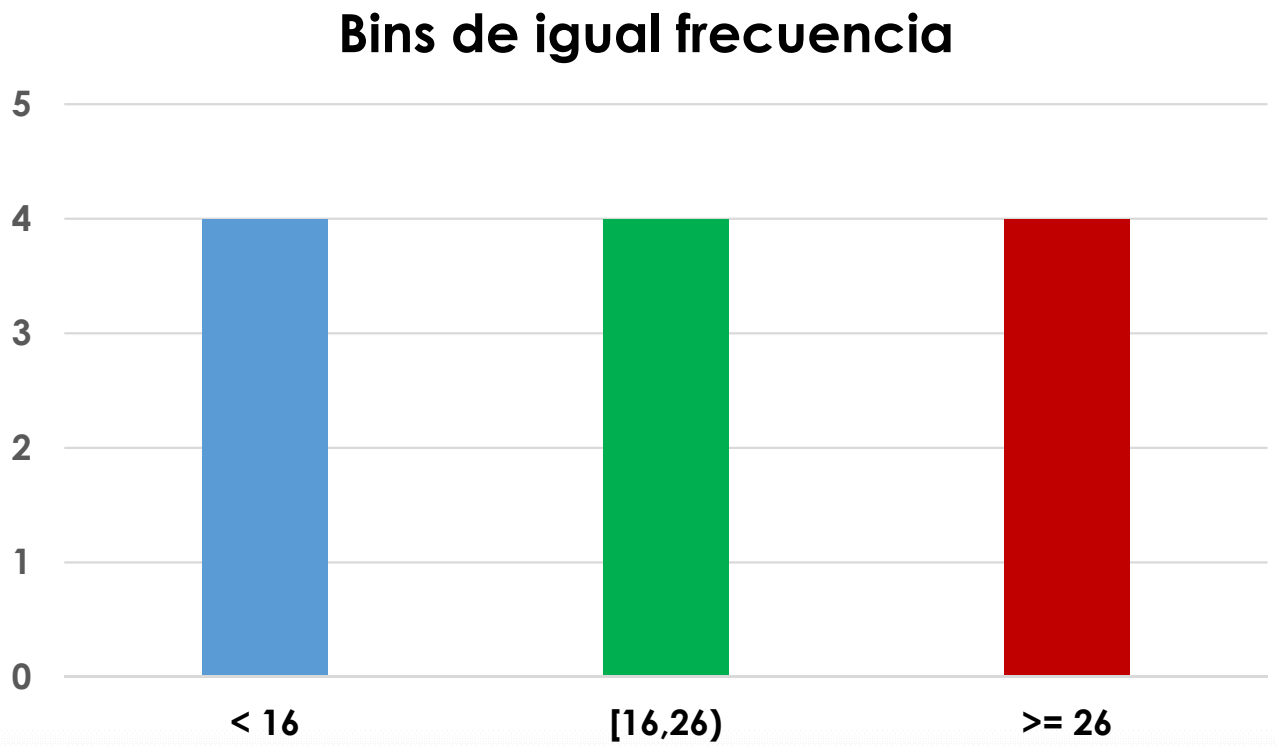
Bins del mismo tamaño	
Bin 1:	4, 8, 9, 15
Bin 2:	21, 21, 24, 25
Bin 3:	26, 28, 29, 34

Suavizado por medias	
Bin 1:	9, 9, 9, 9
Bin 2:	23, 23, 23, 23
Bin 3:	26, 26, 26, 26

Suavizado por valores extremos	
Bin 1:	4, 4, 4, 15
Bin 2:	21, 21, 25, 25
Bin 3:	26, 26, 26, 34

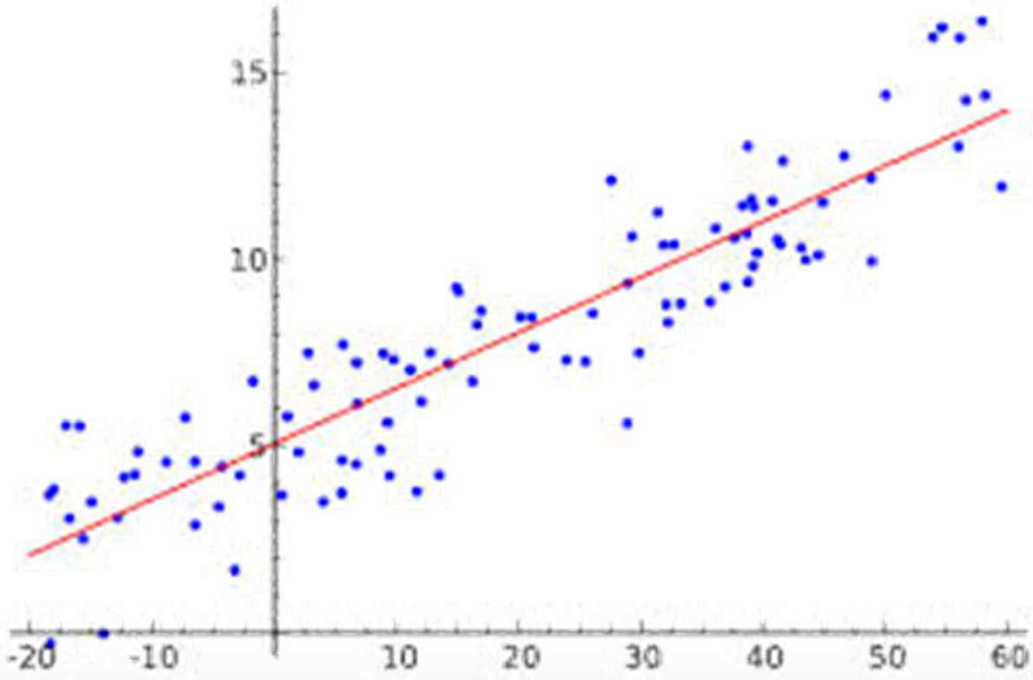
…Discretización simple: *Binning*

Bins del mismo tamaño	
Bin 1:	4, 8, 9, 15
Bin 2:	21, 21, 24, 25
Bin 3:	26, 28, 29, 34



Análisis de regresión lineal

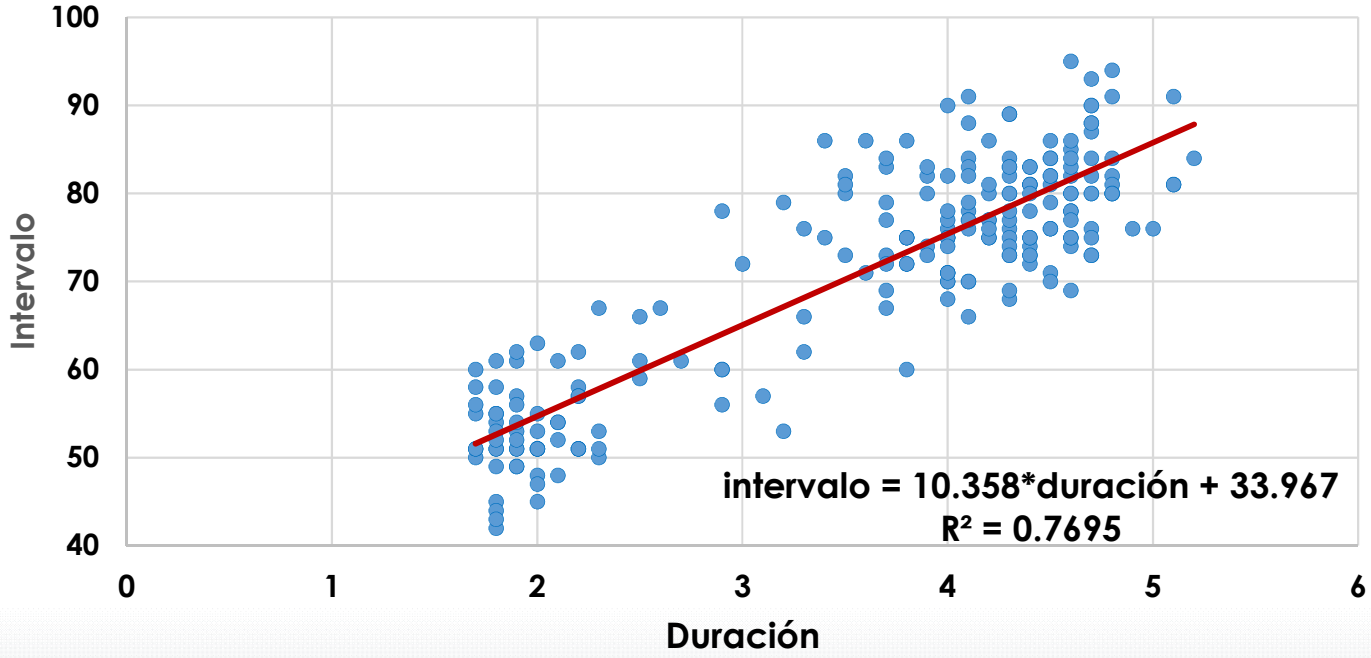
- **Regresión Lineal**
  - Técnica que ajusta los datos a una función.
  - Involucra encontrar la **mejor línea** que ajusta dos atributos (o variables) de manera que un atributo puede ser utilizado para predecir el otro.
  - La **regresión múltiple** es una extensión de la regresión lineal en donde se involucran más de dos atributos que se ajustan a una superficie multidimensional.



…Análisis de regresión

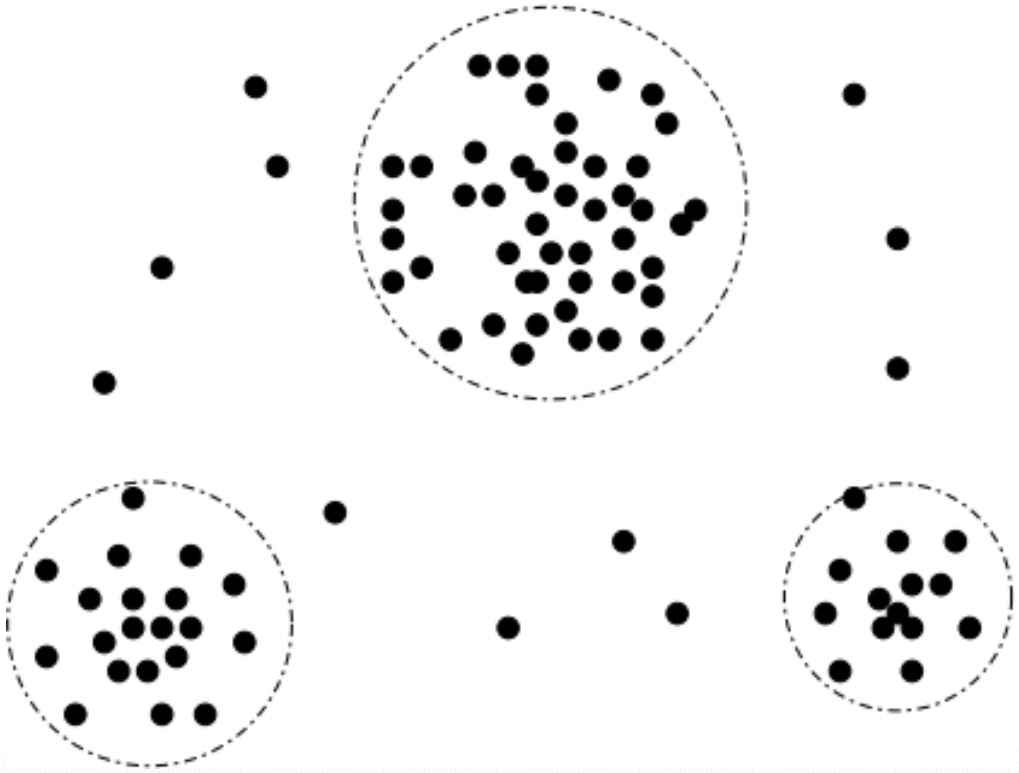
Duración	Intervalo	Predicción	Error
4.4	78	79.5422	1.98%
3.9	74	74.3632	0.49%
4	68	75.399	10.88%
4	76	75.399	-0.79%
3.5	80	70.22	-12.23%
4.1	84	76.4348	-9.01%
2.3	50	57.7904	15.58%
…	…	…	…

Erupción geiser Old Faitful



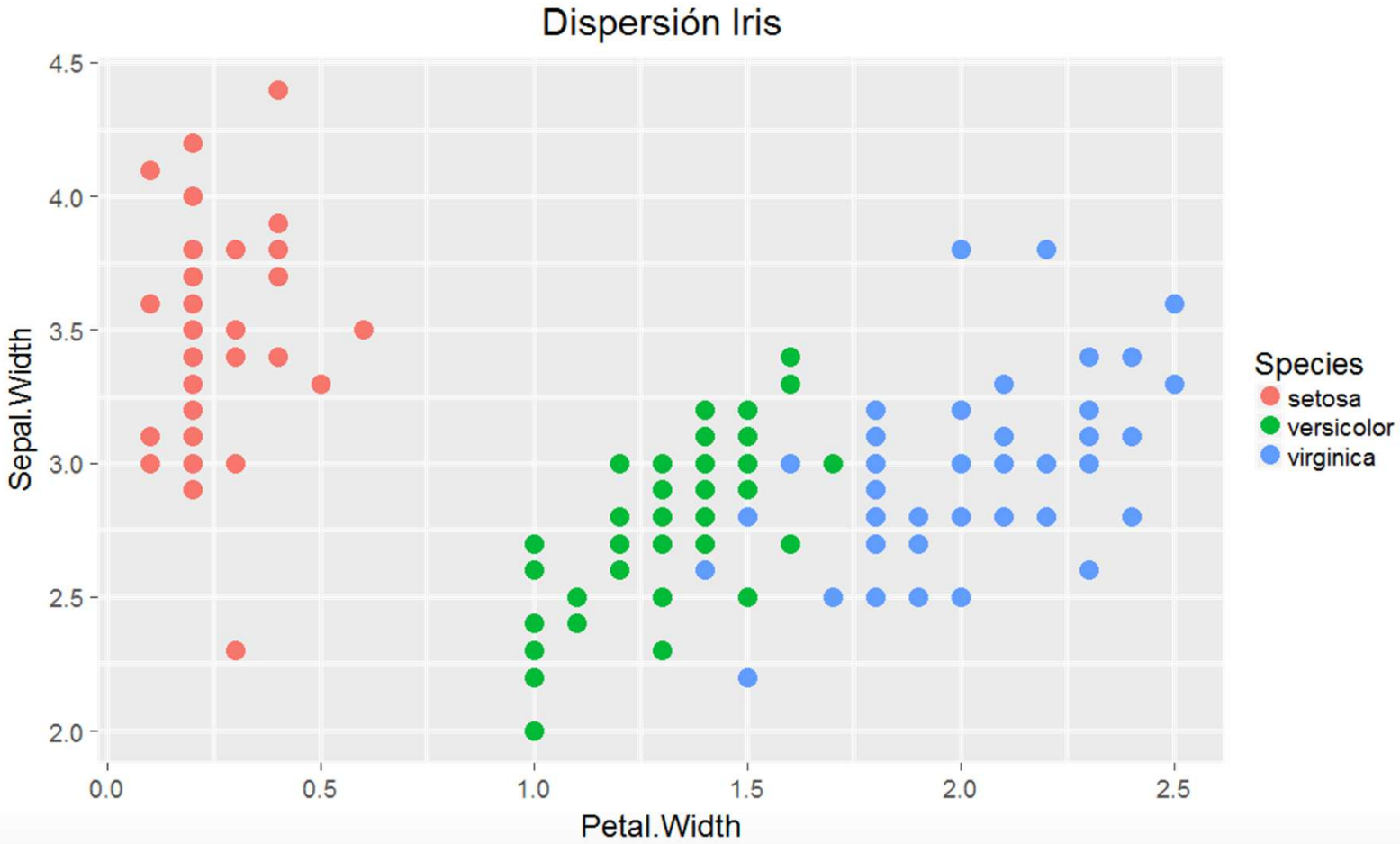
- **Análisis de datos atípicos (Outliers)**

  Los valores atípicos pueden detectarse utilizando técnicas de agrupación (clúster), donde los valores son organizados en grupos de valores similares. Intuitivamente los valores que caigan fuera de esos grupos se consideran como valores atípicos.



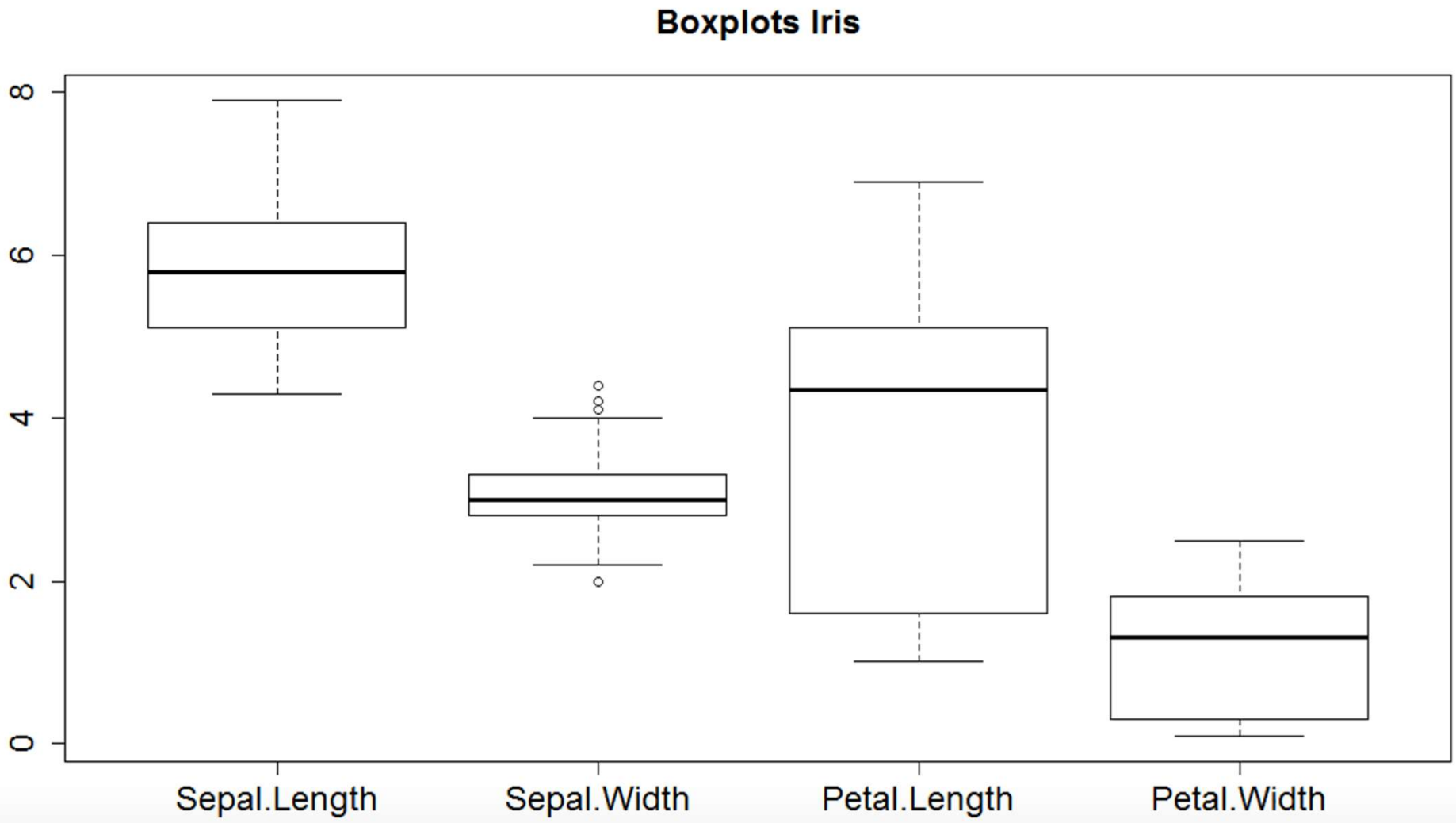
...Outliers

```
> data("iris")
> attach(iris)
> library(ggplot2)
> qplot(Petal.Width,Sepal.Width, data = iris, colour = Species, size = I(4))
```



…Outliers

```
> boxplot(x=iris[,1:4],main="Boxplots Iris",cex.lab=1.5,
          cex.axis=1.5, cex.main=1.5, cex.sub=1.5)
```



Integración de datos

Objetivo: *combinar datos desde distintas fuentes de datos en un almacén coherente, involucra:*

- **Integración de esquemas**
  - Integrar metadatos de distintas fuentes.
  - Problema de identificación de entidades
- **Detectar y resolver los conflictos de valores en los datos**
  - Por la misma razón que las entidades en la vida real, los valores de atributos de diferentes fuentes son diferentes.
  - Diferentes representaciones, diferentes escalas, etc.

Datos redundantes en la integración

- Cuando integran datos provenientes de múltiples fuentes es altamente probable que se presenten **redundancia** en los datos:
  - *El mismo atributo tiene diferentes nombres en diferentes bases de datos.*
  - *Un atributo puede ser derivado (calculado) en otra tabla, p.e. ingresos anuales*
- Los datos redundantes pueden ser detectados a partir de un **análisis de correlación**.
- **Conclusión:**
  - Una integración de datos cuidadosa puede ayudarnos a reducir (en algunos casos evitar) redundancias e inconsistencias en el conjunto de datos resultante.
  - También puede ayudar a mejorar la exactitud y velocidad en los procesos posteriores de minería de datos.
  - Tanto la semántica heterogénea y la estructura de los datos presentan grandes retos para la integración.

Reducción de datos

- **Objetivo:**

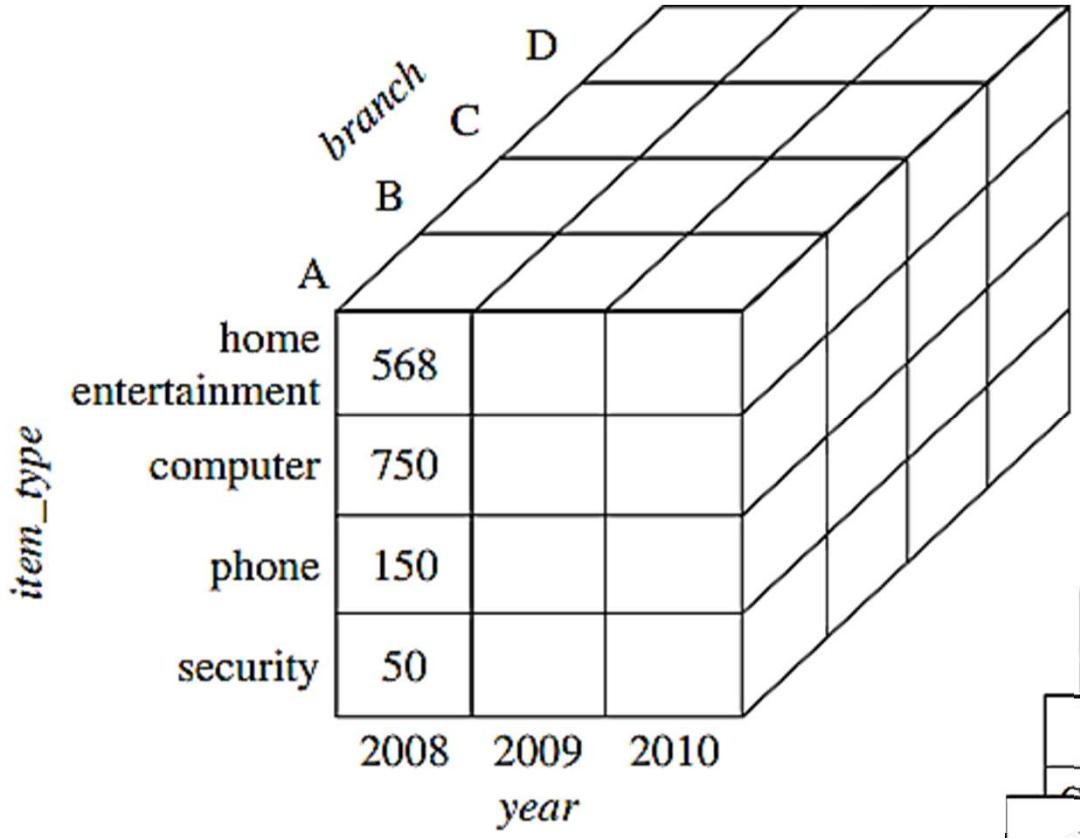
  *Permite obtener una **representación reducida** de un conjunto de datos, que es mucho **menor en volumen**, pero que mantiene la **integridad original** de los datos y que produce **los mismos** (o casi iguales) resultados analíticos.*

- Las estrategias de reducción de datos incluyen:
  - ❑ **Agregaciones en cubos de datos**
  - ❑ **Reducción de dimensionalidad**
  - ❑ **Reducción numerosidad**
  - ❑ **Discretización y la creación de jerarquías concepto.**

Agregaciones en cubos de datos

- Los cubos de datos **describen información** con respecto a varias dimensiones:
  - El contenido de los cubos son **datos de resumen** (*agregaciones calculadas con anticipación*), descritos según aspectos interesantes para la empresa.
- Las agregaciones son el **nivel más bajo** de un cubo de datos:
  - Los datos son agregados para una entidad individual de interés.
  - por ejemplo, el número de llamadas que ha realizado un cliente o las compras que ha hecho.
- **Múltiples niveles** de agregación en cubos de datos
  - Proporcionados por **las jerarquías de concepto**, las cuales permiten explorar los datos desde distintos niveles de abstracción.
- Se debe utiliza la **representación más pequeña** la cual sea suficiente para resolver una tarea en cuestión:
  - Las consultas relativas a la información agregada deben responderse mediante cubo de datos, siempre que sea posible.

…Agregaciones en cubos de datos



Year 2010

Quarter	Sales

Year 2009

Year 2008

Quarter	Sales
Q1	$224,000
Q2	$408,000
Q3	$350,000
Q4	$586,000

→

Year	Sales
2008	$1,568,000
2009	$2,356,000
2010	$3,594,000

Reducción dimensional

- Es el proceso de reducir el número de variables aleatorias a considerar en el análisis.
- **Análisis de correlación y covarianza**
- **Análisis de componentes principales**
  - Proyecta los datos originales en un espacio pequeño.
- **Selección de características** (selección de un subconjunto de atributos):
  - Seleccionar un conjunto mínimo de características tales que la distribución de probabilidad de los valores de las diferentes clases, sea tan cercana como sea posible a la distribución original.
  - Reduce el número de patrones → más fácil de entender.
- **Métodos heurísticos** (número exponencial de opciones).
  - Selección hacia adelante, eliminación hacia atrás o combinación.
  - Inducción de árboles de decisión

Análisis de correlación

- Un atributo (*p.e. ingresos anuales*) puede ser redundante si éste puede obtenerse de otro atributo o conjunto de atributos.
- Inconsistencias en la dimensión de un atributo o su nombre pueden resultar en redundancia en el conjunto resultante.
- Algunos aspectos de redundancia pueden ser detectados a través de un **análisis de correlación**:
  - *A través de este análisis, dados dos atributos, se puede medir que tan fuertemente relacionados se encuentran, basados en los datos disponibles.*
  - *Para datos nominales se puede utilizar la prueba* ***Ji-Cuadrada (χ^2)***.
  - *Para atributos numéricos se puede utilizar el* ***coeficiente de correlación*** *y la* ***covarianza*** *(indican una medida de cómo los valores de los atributos varían unos de otros)*.

Análisis de correlación (χ^2)

- Una relación de correlación entre dos atributos **A y B** (*nominales*), pueden ser descubiertos a través de una prueba χ^2:
  - *Vamos a suponer que **A** tiene **c** distintos valores, llamados $a_1, a_2, \ldots, a_c$. **B** tiene **r** distintos valores llamados $b_1, b_2, \ldots, b_r$. Los datos de las tuplas descritas por **A** y **B** se pueden mostrar como una **tabla de contingencia**, con los **c** valores de **A** colocados en **columnas** y los **r** valores de **B** colocados como **renglones**.*
  - **(A_i, B_j)** ***denota un evento conjunto donde el atributo A toma un valor a_i y el atributo B toma el valor b_j. Todos y cada uno de los posibles eventos tiene su propia celda en la tabla:***

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{\left(o_{ij} - e_{ij}\right)^2}{e_{ij}} \quad \text{donde} \quad \begin{array}{l} o_{ij} \text{ es la frecuencia observada} \\ e_{ij} \text{ es la frecuencia esperada} \end{array}$$

$$e_{ij} = \frac{conteo\left(A = a_i\right) \times conteo\left(B = b_j\right)}{n}$$

  - ***A partir de este análisis se prueba la hipótesis de que A y B son independientes (no hay correlación).***
  - ***Se basa en un nivel de significa con* (r-1) x (c-1) *grados de libertad.***

…Análisis de correlación (χ^2)

Supongamos que un grupo de **236 estudiantes** fue examinado. Se anotó información sobre sus **hábitos de fumar** *(en exceso, regularmente, ocasionalmente y no fuma)*. También se les preguntó sobre su **nivel de ejercicio** *(frecuentemente, en ocasiones, nunca)*.

- *Se tienen dos atributos, fumar y nivel de ejercicio. La frecuencia observada de cada evento conjunto posible se muestra en la tabla*:

	Frecuentemente	En ocasiones	Nunca	Total
En exceso	7	3	1	**11**
Regularmente	9	7	1	**17**
Ocasionalmente	12	4	3	**19**
Nunca	87	84	18	**189**
Total	**115**	**98**	**23**	**236**

- *Calculemos las **frecuencias esperadas**. Por ejemplo para el par **(en exceso, frecuentemente)**:*

$$e_{11} = \frac{conteo(\text{En exceso}) \times conteo(\text{Frecuentemente})}{\text{total personas}} = \frac{11 \times 115}{236} = 5.36$$

…Análisis de correlación (χ^2)

- *Los números entre paréntesis corresponden a las frecuencias esperadas:*

	Frecuentemente	En ocasiones	Nunca	Total
En exceso	7 **(5.36)**	3 **(4.57)**	1 **(1.07)**	**11**
Regularmente	9 **(8.28)**	7 **(7.06)**	1 **(1.66)**	**17**
Ocasionalmente	12 **(9.26)**	4 **(7.89)**	3 **(1.85)**	**19**
Nunca	87 **(92.10)**	84 **(78.48)**	18 **(18.42)**	**189**
Total	**115**	**98**	**23**	**236**

- El cálculo para χ^2 será:

$$\chi^2 = \frac{(7-5.36)^2}{5.36} + \frac{(3-4.57)^2}{4.57} + \frac{(1-1.07)^2}{1.07} + \frac{(9-8.28)^2}{8.28} + \frac{(7-7.06)^2}{7.06} +$$
$$\frac{(1-1.66)^2}{1.66} + \frac{(12-9.26)^2}{9.26} + \frac{(4-7.89)^2}{7.89} + \frac{(3-1.85)^2}{1.85} + \frac{(87-92.10)^2}{92.10} +$$
$$\frac{(84-78.48)^2}{78.48} + \frac{(18-18.42)^2}{18.42}$$

$$\chi^2 = 0.5 + 0.54 + 0 + 0.06 + 0 + 0.26 + 0.81 + 1.92 + 0.71 + 0.28 + 0.39 + 0.01 = 5.49$$

…Análisis de correlación (χ^2)

- Para una tabla de **4x3**, los grados de libertad son **(4-1)(3-1) = 6**.
- Para **4 grados de libertad**, el valor necesario para **rechazar la hipótesis** con un nivel de significancia de **0.001** es:

	p										
g	0.001	0.025	0.05	0.1	0.25	0.5	0.75	0.9	0.95	0.975	0.999
1	10.827	5.024	3.841	2.706	1.323	0.455	0.102	0.016	0.004	0.001	0
2	13.815	7.378	5.991	4.605	2.773	1.386	0.575	0.211	0.103	0.051	0.002
3	16.266	9.348	7.815	6.251	4.108	2.366	1.213	0.584	0.352	0.216	0.024
4	18.466	11.143	9.488	7.779	5.385	3.357	1.923	1.064	0.711	0.484	0.091
5	[illegible]	12.832	11.07	9.236	6.626	4.351	2.675	1.61	1.145	0.831	0.21
6	22.457	14.449	12.592	10.645	7.841	5.348	3.455	2.204	1.635	1.237	0.381

- Como el valor que se obtuvo (**5.49**) esta por debajo del indicado en la tabla, la hipótesis comprueba: **el hábito de fumar de los estudiantes es independiente del nivel de ejercicio**.

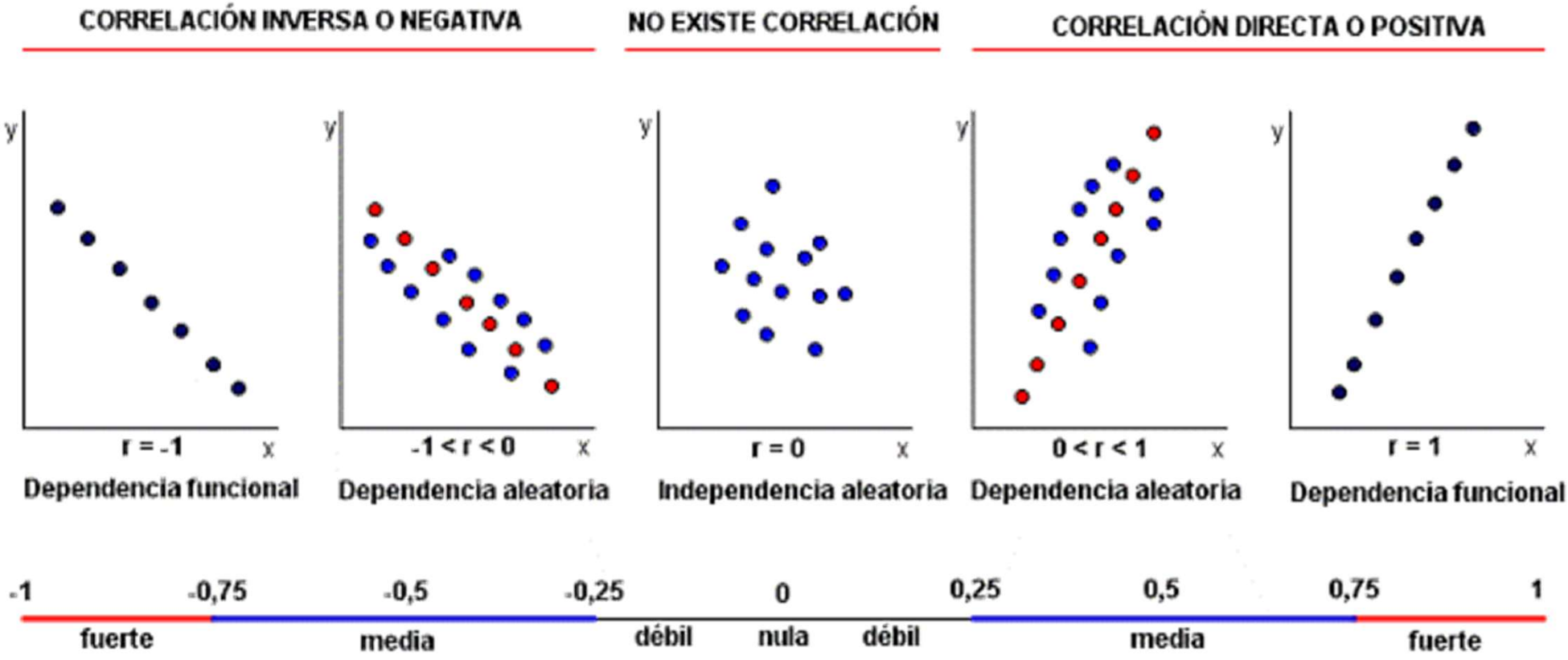
Coeficiente de correlación

- El **coeficiente de correlación de Pearson**, pensado para variables cuantitativas, es un índice que mide el grado de **covariación** entre distintas variables relacionadas linealmente.
- Se define de la siguiente forma:

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{N\sum x_i^2 - \left(\sum x_i\right)^2}\sqrt{N\sum y_i^2 - \left(\sum y_i\right)^2}}$$

- A diferencia de la covarianza, la **correlación de Pearson** es independiente de la escala de medida de las variables.
- Se basa en la prueba de **t de Student**, para determinar el **nivel de significancia** del coeficiente (esto es, **aprobar o rechazar** la hipótesis nula).

…Coeficiente de correlación

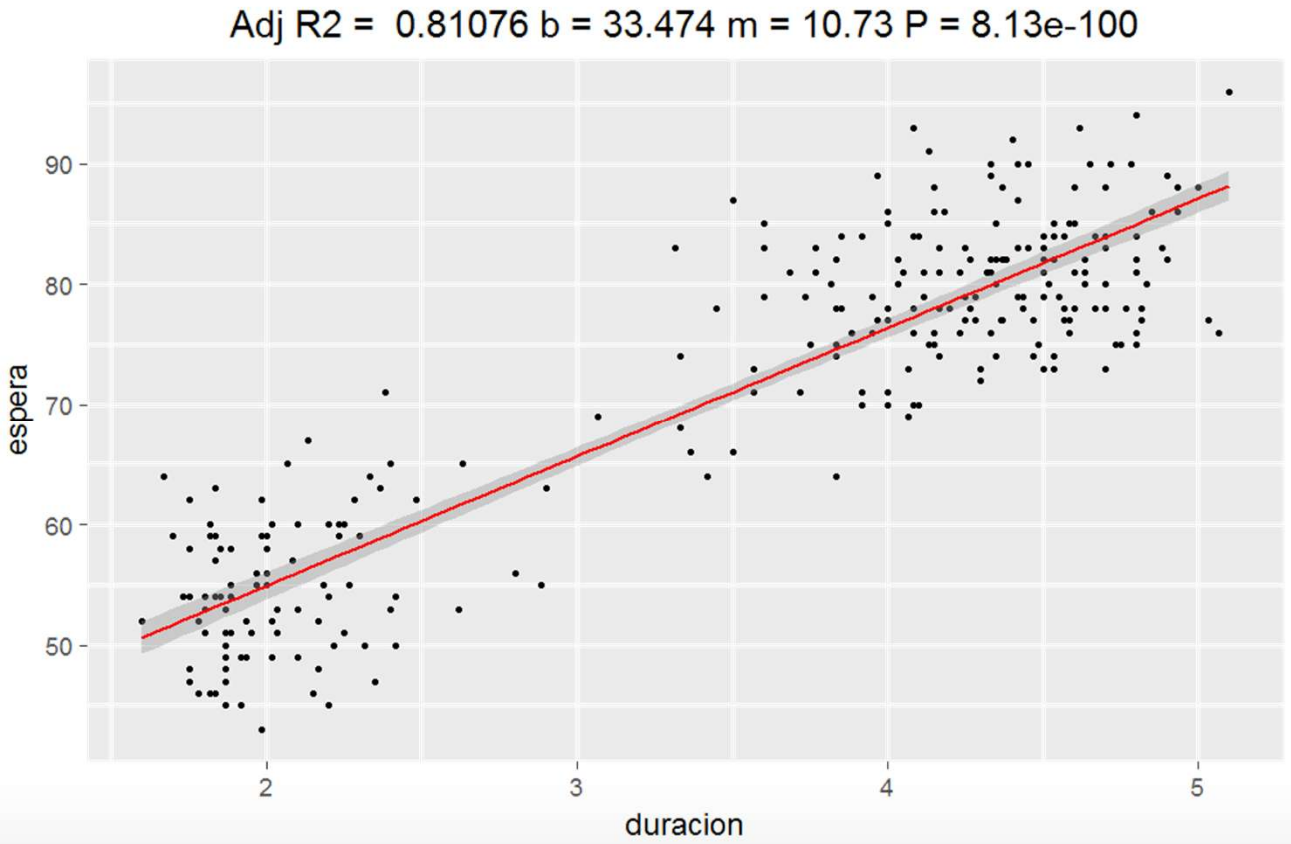


Es importante notar que **correlación** no significa **causalidad**, es decir, si A y B están correlacionadas, no significa necesariamente que A implica B o B implica A.

…Coeficiente de correlación

```
#Información del Geiser de Yellowstone
> head(faithful)
> attach(faithful)
#Obtenemos la duración y la espera
> duracion = eruptions
> espera = waiting
> head(cbind(duracion, espera))
> fit <- lm(espera ~ duracion)
#Prueba de correlación
> cor(duracion, espera)
```

```
> cor(duracion, espera)
[1] 0.9008112
```



Análisis de componentes principales

- Cuando se va a analizar la información de una muestra de datos, lo más frecuente es tomar el mayor número posible de variables.
- Problemas:
  - **Muchas variables → Demasiados coeficientes de correlación**

$$\binom{20}{2} = \frac{20!}{2!(20-2)!} = 190 \text{ coeficientes de correlación}$$

$$\binom{40}{2} = \frac{40!}{2!(40-2)!} = 780 \text{ coeficientes de correlación}$$

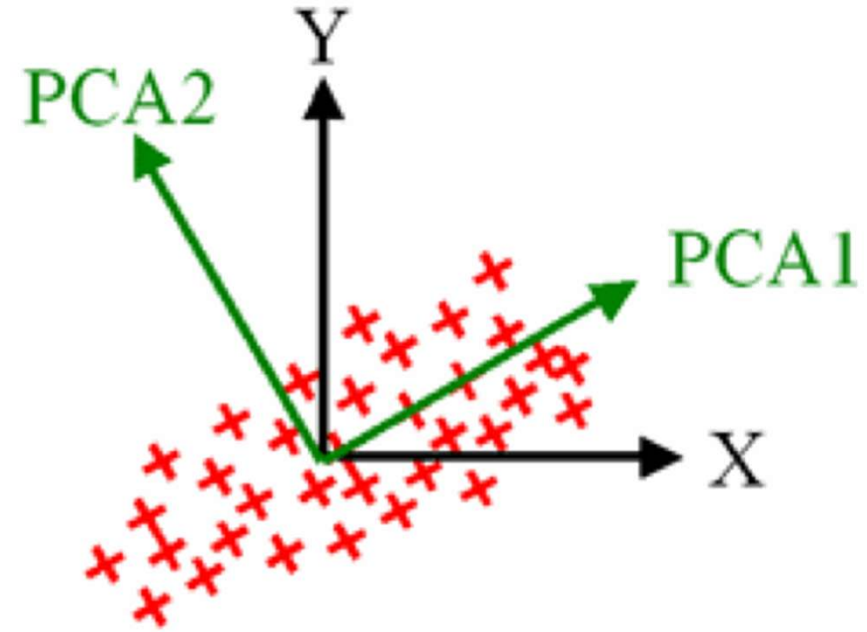
Entre más variables se seleccionen, más difícil será visualizar la correlación

  - **Fuerte correlación → Muchas variables miden lo mismo**

- El **Análisis de Componentes Principales** soluciona esta problemática.

…Análisis de componentes principales

- Se trata de una **técnica multivariada** que se conoce como **Métodos Factoriales**.
  - Tiene como objetivo **resumir** un gran conjunto de datos.
  - Se basa en la **creación de nuevas variables** que son **función lineal** de las variables originales.
  - De esta forma, reduce la dimensión de un conjunto de variables a un conjunto de menor número de variables que permitan mejorar la **interpretabilidad** de los datos.



PCA: Fases

- **Fase 1. Análisis de la Matriz de Correlaciones:**
  - Solo se puede aplicar este análisis si existen altas correlaciones entre las variables (**alta correlación → información redundante**)

- **Fase 2. Selección de factores:**
  - Se busca que el **primer factor** explique la mayor cantidad de variabilidad.
  - El **segundo factor** debe explicar la máxima variabilidad posible que no fue explicada por el primer factor y así sucesivamente.
  - Se toman en cuenta los factores que expliquen el **mayor porcentaje de variabilidad**.
  - La selección del número de factores depende del analista o de las características del problema.

…PCA: Fases

- **Fase 3. Análisis de la matriz factorial:**
  - Una vez seleccionados los componentes principales, se representan en forma de matriz, cada elemento representan los coeficientes factoriales de las variables (***correlaciones entre las variables y los componentes principales***).
  - La matriz tendrá tantas columnas como CP y filas como variables
- **Fase 4. Interpretación de los factores:**
  - Depende enteramente del analista, un factor es más fácil de interpretar si:
    1. **Coeficientes factoriales ente -1 y 1**
    2. **Una variable debe tener coeficientes factoriales elevados solo con un factor.**
    3. **No deben existir factores con coeficientes factoriales similares**

Selección de características

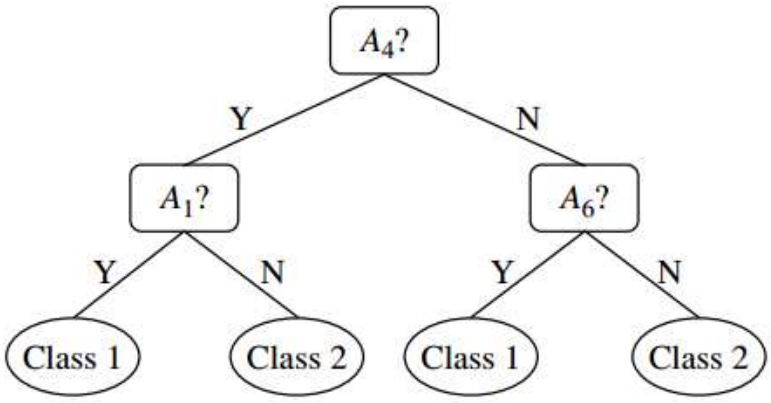
- **Objetivo:** encontrar el **mínimo conjunto de atributos** tal que la distribución de probabilidad resultante de los datos sea **tan cercana como sea posible** a la distribución obtenida de utilizar todos los atributos.
- El **"mejor" subconjunto de atributos**:
  - Para **n atributos**, existen **2^n subconjuntos** posibles *(una búsqueda exhaustiva sería muy costosa*), de manera que se utilizan heurísticas que reducen el espacio de búsqueda.
  - Se pueden utilizar **algoritmos voraces**, los cuales mientras buscan en el espacio de atributos, siempre hacen lo que parece ser ***la mejor opción en el momento***.
  - Hacen una **elección localmente óptima** con la esperanza de que esto conduzca a una solución óptima global.

...Subconjuntos de atributos

- **Selección hacia adelante.**
  - El procedimiento inicia con un conjunto vacío de atributos (conjunto reducido inicial). El **mejor** de los atributos originales se determina y se añade al conjunto reducido. En cada iteración, el mejor de los atributos originales restantes se añade al conjunto.
- **Eliminación hacia atrás.**
  - El procedimiento inicia con el conjunto completo de atributos. En cada paso, se elimina el peor atributo que queda en el conjunto.
- **Selección hacia adelante y eliminación hacia atrás.**
  - En cada paso, el procedimiento selecciona el mejor atributo y elimina el peor de entre los atributos restantes.
- **Árbol de decisión.**
  - Construye un diagrama donde cada nodo interno (no hoja) denota una prueba en un atributo, cada rama corresponde a un resultado de la prueba, y cada nodo hoja denota una predicción . En cada nodo, los algoritmo elige "mejor" de atributos para dividir los datos en clases individuales.

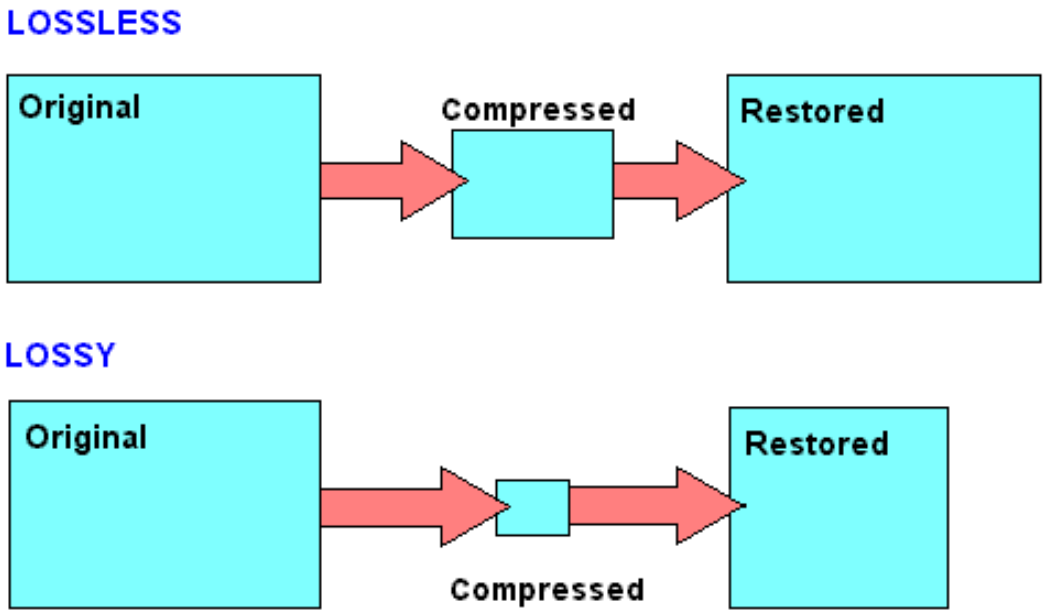
...Subconjuntos de atributos

- Los "mejores" (y "peores") atributos se determinan normalmente utilizando pruebas de **significancia estadística**:
  - Asumen que los atributos son independientes entre sí.
  - Otras medidas de evaluación de atributos que se pueden utilizar son la ganancia de información (utilizada en la construcción de árboles de decisión para la clasificación).

Forward selection	Backward elimination	Decision tree induction
Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ Initial reduced set: $\{\}$ $\Rightarrow \{A_1\}$ $\Rightarrow \{A_1, A_4\}$ => Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ $\Rightarrow \{A_1, A_3, A_4, A_5, A_6\}$ $\Rightarrow \{A_1, A_4, A_5, A_6\}$ => Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$  => Reduced attribute set: $\{A_1, A_4, A_6\}$

Compresión de datos

- **Compresión de datos**, aplica transformaciones para obtener representaciones reducidas de los datos originales.
  - Si los datos originales pueden ser reconstruidos de los datos comprimidos sin pérdida de información se conocen como **algoritmos sin pérdida**.
  - Si por el contrario solo se puede reconstruir una aproximación, entonces se conocen como **algoritmos con pérdida**.



Reducción de numerosidad

- Es una técnica que reemplaza el volumen original de datos por una forma alternativa de representación.
  - **Métodos paramétricos:**

    Supone que los datos se ajusta a algún modelo y estima los parámetros del modelo, almacenar sólo los parámetros, y descartar los datos (excepto posibles valores atípicos)

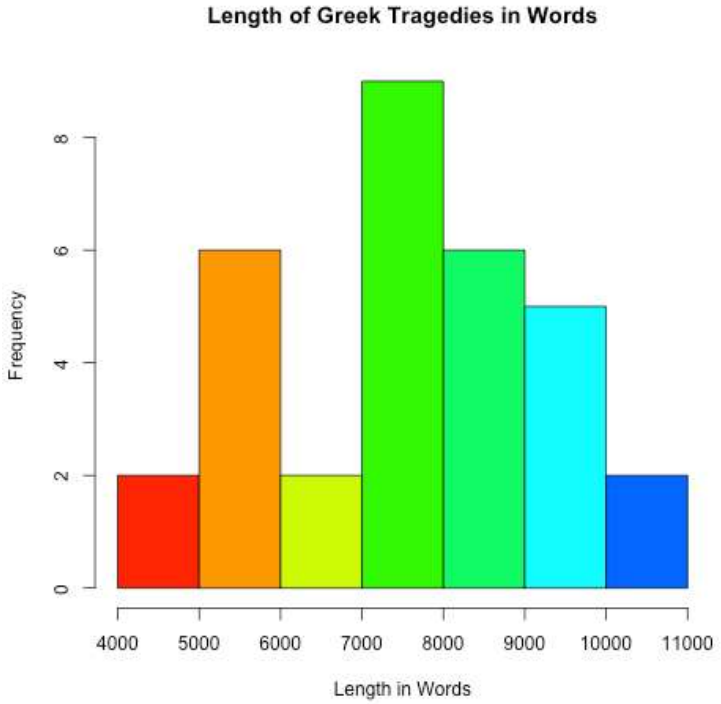
  - **Métodos no paramétricos:**

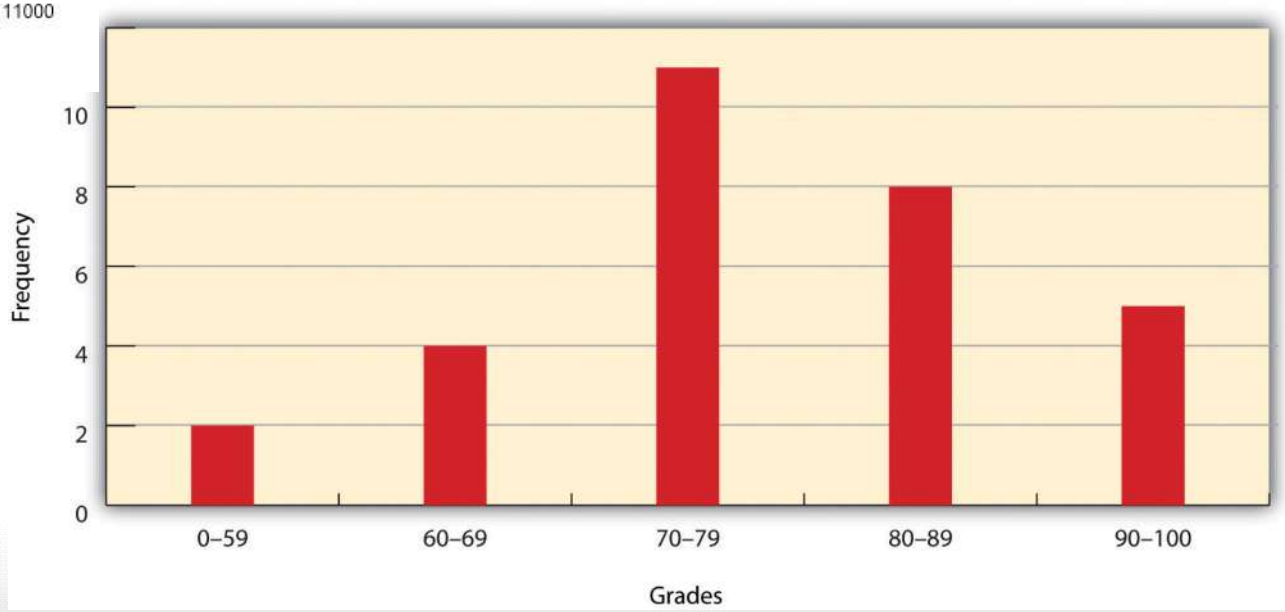
    No asume ningún modelo, incluye: histogramas, agrupamiento y el muestreo.

Histogramas

- Una técnica popular de reducción de datos además de ser útiles para detectar valores atípicos (la columna asociada a ellos aparecerá alejada del resto y posiblemente con poca frecuencia).
- Crea una partición sobre un atributo, la cual puede hacerse de distintas formas: cada valor del atributo define un subconjunto o bien, se crean intervalos sobre el atributo.
- Los subconjuntos resultantes de la partición se denominan contenedores o "bins".
- Utilizan reglas de partición, las más comunes son:
  - **Igual ancho:** En este histograma, el ancho de cada contenedor es uniforme.
  - **Igual frecuencia:** En este histograma, se crean los cubos de modo que, más o menos, la frecuencia de cada cubo es constante (es decir, cada cubo contiene aproximadamente el mismo número de muestras de datos contiguos).
- Los histogramas son altamente eficaces en la aproximación de datos escasos y densos, así como datos altamente asimétricos y uniformes.

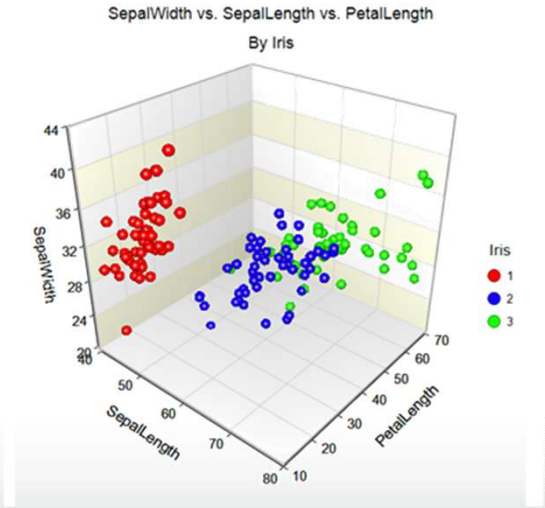
…Histogramas





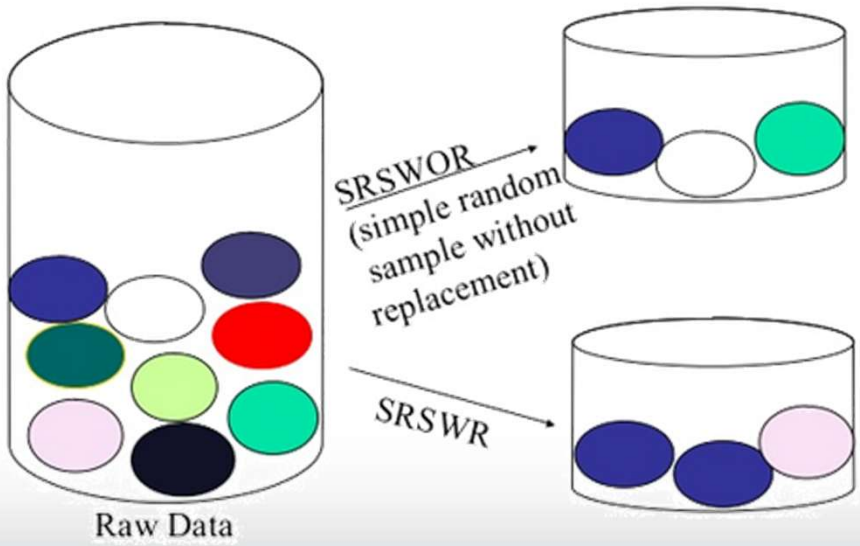
Agrupación

- Consideran tuplas de datos como objetos y particionan los objetos en conglomerados , por lo que los objetos dentro de un grupo son "similares" entre sí y "diferentes" a los objetos de otros grupos.
- La similitud se define comúnmente en términos de ***qué tan cerca*** se encuentran los objetos están en el espacio (***función de distancia***).
- La "calidad" de un grupo puede ser representado por su diámetro (distancia máxima entre dos objetos en el clúster).
- La **distancia del centroide** es una medida alternativa de calidad del grupo y se define como la distancia media de cada objeto de clúster desde su centroide (***punto medio en el espacio para el clúster***).

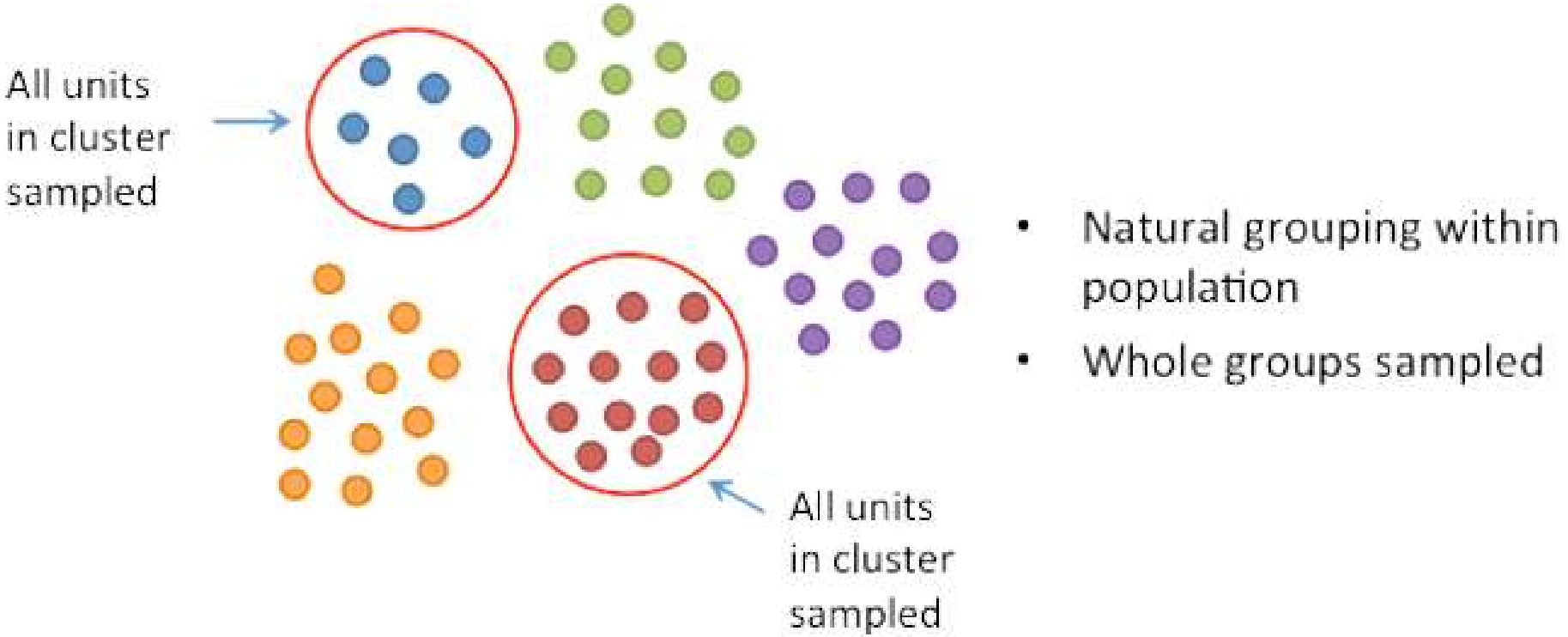


Muestreo

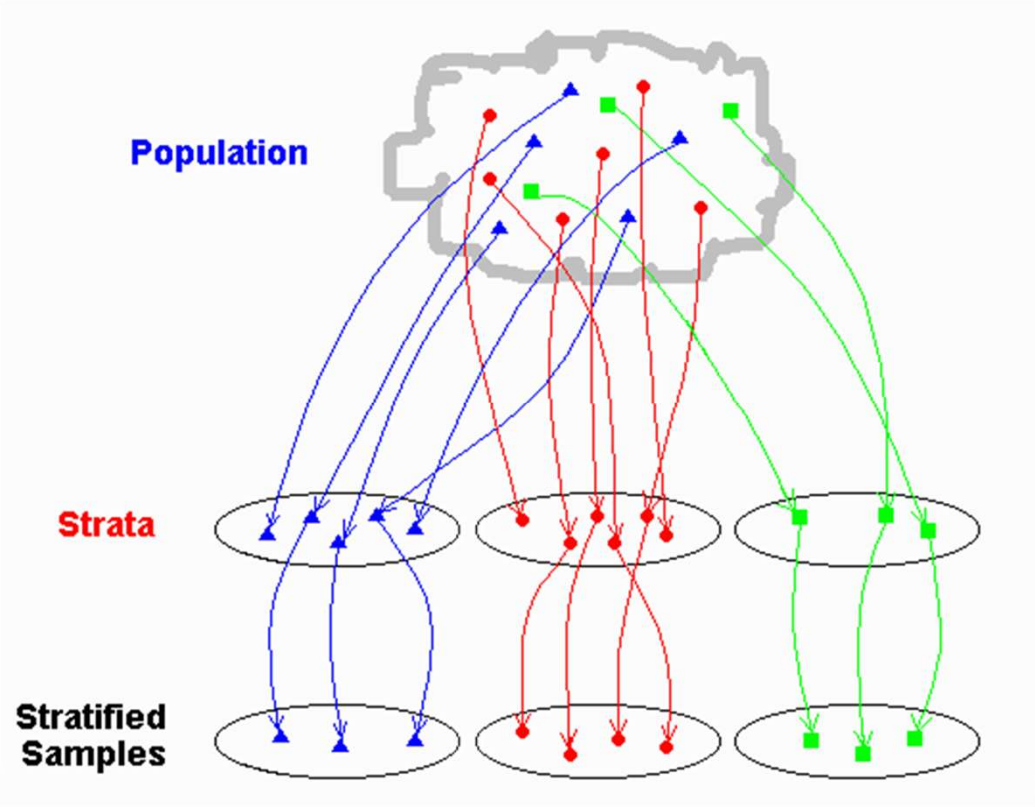
- Este enfoque permite crear un conjunto de datos de menor dimensión, creado con datos tomados al azar del original.
- Se busca reducir el conjunto de datos obteniendo los mismos resultados.
- **Muestreo aleatorio simple sin reemplazo de tamaño s**.
  - Se crea tomando s tuplas (s < N), donde todas las tuplas son misma probabilidad de ser seleccionadas.
- **Muestreo aleatorio simple con reemplazo de tamaño s.**
  - Similar al muestreo anterior, excepto que cada vez que una tupla se extrae de D, se registra y se reemplaza. Es decir, después de que se extrae una tupla, se coloca de nuevo en D de modo que puede extraerse de nuevo.



- **Muestreo de clúster:**
  - Si las tuplas en D se agrupan en M grupos mutuamente excluyentes entonces un muestreo aleatorio simple se puede obtener, en donde s <M.



- **Muestreo estratificado.**
  - Si D se divide en partes **mutuamente excluyentes** (estratos), una muestra estratificada de D se genera mediante la obtención de una muestra aleatoria simple en cada estrato. Esto ayuda a garantizar una muestra representativa, especialmente cuando los datos son asimétricos.



…Muestreo

- Una ventaja de muestreo para la reducción de datos es que el costo de la obtención de una muestra es proporcional al tamaño de la muestra y no del conjunto de datos.
- Cuando se aplica a la reducción de datos, el muestreo es más comúnmente utilizado para estimar la respuesta a una consulta de agregaciones.
- Es posible (utilizando el teorema del límite central) determinar un tamaño de muestra suficiente para la estimación de una función dada dentro de un grado de error especificado.
- El muestreo es una elección natural para el perfeccionamiento progresivo de un conjunto de datos reducido. Dicho conjunto se puede refinar aún más por el simple aumento del tamaño de la muestra.

Transformación de datos

- Con esta técnica, los datos son transformados o consolidados para que los resultados del proceso de minería de datos sean más eficientes y los patrones encontrados, más fáciles de entender:
  - **Suavizado:** técnica que para eliminar el ruido de los datos (binning, regresión y agrupación)
  - **Construcción de atributos(*o características*)**: los nuevos atributos se construyen y se agregan a partir del conjunto dado de atributos para ayudar al proceso de minería .
  - **Agregación:** se aplican operaciones de resumen de datos. Este paso se utiliza típicamente en la construcción de un cubo para el análisis de datos en múltiples niveles de abstracción.
  - **Normalización:** los datos de los atributos se escalan de manera que caiga dentro de un rango menor **(-1.0 a 1.0 o de 0.0 a 1.0**).
  - **Generalización.** Jerarquías de concepto.

…Transformación de datos

- **Discretización:** los valores de un atributo numérico (p.e. la edad) son reemplazadas por etiquetas de intervalo (p.e. 0-10, 11-20, etc.) o etiquetas conceptuales (p.e. jóvenes, adultos, etc.). Las etiquetas, a su vez se pueden organizar de forma recursiva en conceptos de alto nivel, lo que resulta en una jerarquía de concepto para el atributo numérico.
- **Jerarquía de Concepto para datos nominales:** algunos los atributos se puede generalizar a conceptos de alto nivel. Muchas jerarquías de atributos nominales están implícitos en el esquema de base de datos y pueden ser definidos de forma automática en el nivel de definición de esquema.

Normalización

- Las unidades de medida pueden afectar el análisis de datos. En general, expresar un atributo en unidades más pequeñas conducirá a un mayor rango para ese atributo, y por lo tanto tienden a dar tal atributo mayor efecto o peso.
- La normalización intenta dar a todos los atributos un peso igual.
- Existen varios métodos para efectuar normalización:
  - **Normalización min-max**: realiza una transformación lineal en los datos originales. Supongamos que $\mathbf{min_A}$ y $\mathbf{max_A}$ son los valores mínimo y máximo de un atributo **A**, esta normalización mapea un valor $\mathbf{v_i}$ en un valor $\mathbf{v_i'}$ en el rango $\mathbf{(nuevo\_min_A, nuevo\_max_A)}$

$$v_i' = \frac{v_i - \min_A}{\max_A - \min_A}\left(nuevo\_\max_A - nuevo\_\min_A\right) + nuevo\_\min_A$$

Se preservan las relaciones entre los datos originales y no permite que valores fuera del conjunto original caigan en el nuevo conjunto.

…Normalización

- **Normalización z-score**: los valores para un atributo A, son normalizados basados en la media y la desviación estándar de A:

$$v_i' = \frac{v_i - \overline{A}}{\sigma_A}$$

Este método de normalización es bastante útil cuando los valores mínimo y máximo del atributo A son desconocidos o bien cuando los outliers dominan la normalización min-max.

Una variación de esta normalización reemplaza la deviación estándar por la desviación media absoluta de A (s_A):

$$s_A = \frac{1}{n}\left(\left|v_1 - \overline{A}\right| + \left|v_2 - \overline{A}\right| + \cdots + \left|v_n - \overline{A}\right|\right) \quad \Rightarrow \quad v_i' = \frac{v_i - \overline{A}}{s_A}$$

Esta normalización es más robusta para los outliers ya que las deviaciones de la media no se elevan al cuadrado y el efecto de los outliers se reduce.

Normalización

- **Normalización por escala decimal**: normaliza moviendo el punto decimal de los valores de atributo A. El número de puntos decimales movido depende del valor máximo absoluto de A**:**

$$v_i' = \frac{v_i}{10^j}$$

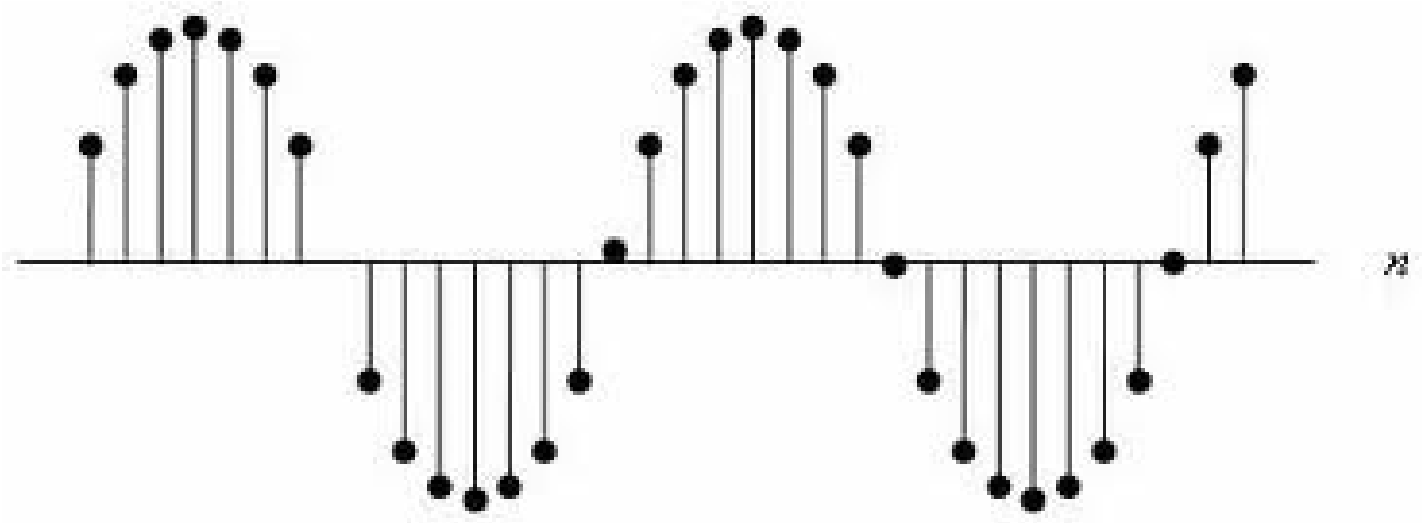
Donde j es el menos entero tal que $\max(|v_i'|) < 1$

Discretización

- Tiene el objetivo de reducir el número de valores de un atributo continuo, dividiendo el rango del atributo en intervalos. Los intervalos etiquetados pueden utilizarse para reemplazar los valores de datos reales.
- Las técnicas de discretización se pueden clasificar en función de cómo se realiza la tarea, por ejemplo si se utiliza la información de clase o en qué dirección procede (arriba hacia abajo vs abajo hacia arriba).
  - Si el proceso de discretización utiliza información de clase, entonces decimos que es discretización **supervisada** (histogramas). De lo contrario, es **sin supervisión**.
  - Si el proceso se inicia mediante la búsqueda un o unos pocos puntos (llamados puntos de división) para dividir todo el rango de atributo, y luego se repite esta forma recursiva en los intervalos resultantes, se denomina **discretización de arriba hacia abajo** o de división (árboles de decisión)

…Discretización

- La **discretización abajo hacia arriba** o mezclado, que parte de considerar todos los valores continuos como potenciales puntos de división, elimina algunos por la mezcla de los valores de vecinos para formar intervalos, y luego se aplica de forma recursiva este proceso para los intervalos resultantes.

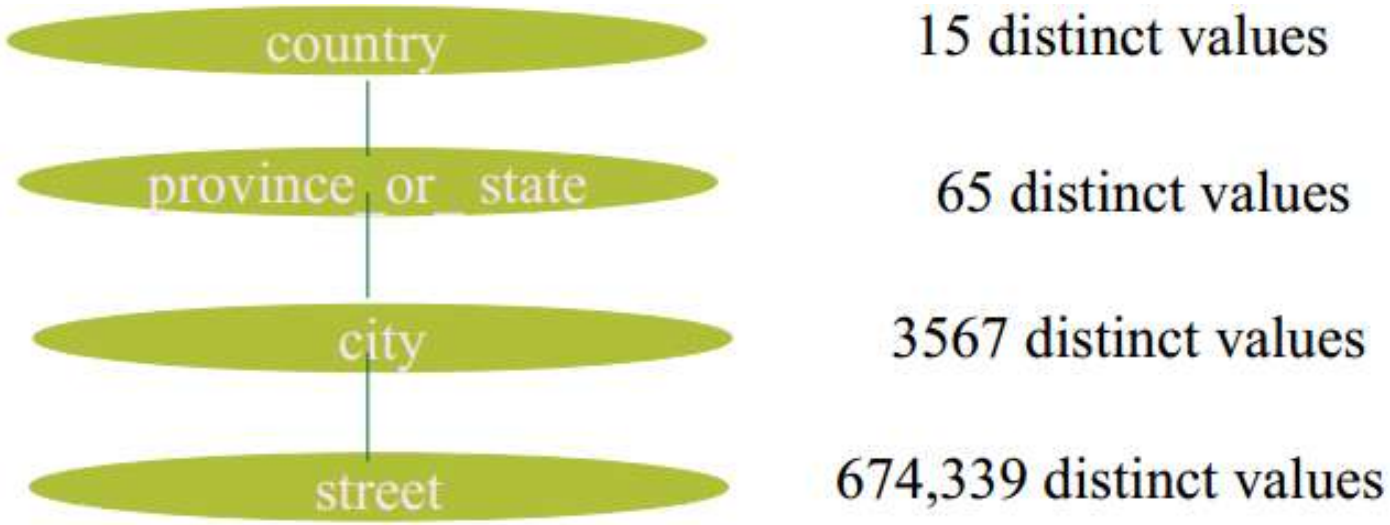


Jerarquía de conceptos

- Tiene como objetivo reducir los datos mediante la recopilación y la sustitución de los conceptos de bajo nivel (valores numéricos para el atributo edad) por los conceptos de nivel superior (p.e. jóvenes, adultos, adultos mayores).
- Los atributos nominales tienen un número finito (pero posiblemente grande) de valores distintos, sin ningún orden entre los valores (p.e. ubicación, categoría laboral, tipo de artículo etc.).
- La definición de jerarquías de conceptos puede ser una tarea tediosa y consume mucho tiempo para un usuario e incluso un experto.
  - Una ventaja que se tiene es que muchas jerarquías están implícitas en el esquema de la BD y pueden ser definidas de forma automática cuando se define el esquema.
- Esta técnica se utiliza para transformar los datos en múltiples niveles de granularidad.

...Jerarquía de conceptos

- Una jerarquía concepto se puede generar automáticamente en función del número de valores distintos por cada atributo en el conjunto de atributo dado.
  - El atributo con los valores más diferenciados se sitúa en el nivel de jerarquía más baja.
  - Cuanto menor es el número de valores distintos de un atributo tiene, cuanto más alto se encuentra en la jerarquía concepto generado.



Conclusión

Una buena preparación de datos es clave para producir modelos válidos y fiables