# Unidad 5

Exploración y selección

Modelizado de Minería de Datos - Q22025

# Introducción

Una vez los datos están recopilados, integrados y “limpios”, es necesario, además, realizar un reconocimiento o análisis exploratorio (EDA) de los datos con el objetivo de **conocerlos mejor**.

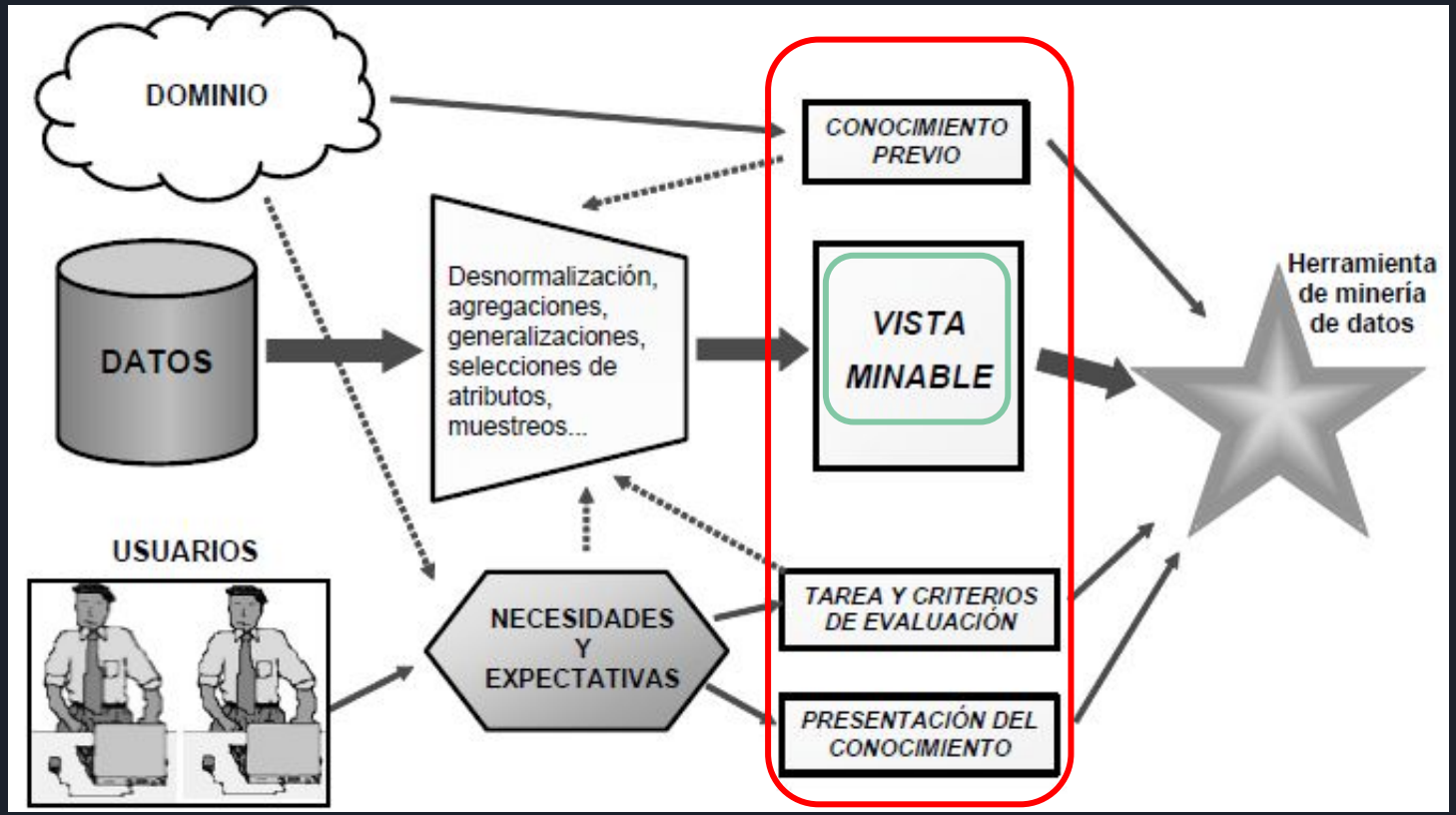
El output estas técnicas se lo suele conocer como “**vista minable**” la cual, además, cuenta con “instrucciones” sobre qué datos trabajar, qué tarea realizar y de qué manera obtener el conocimiento.

# El contexto de la vista minable

- ¿Qué parte de los datos es pertinente analizar?
- ¿Qué tipo de conocimiento se desea extraer y cómo se debe presentar?
- ¿Qué conocimiento puede ser válido, novedoso e interesante?
- ¿Qué conocimiento previo me hace falta para realizar esta tarea?

Incluso conociendo los datos y el dominio del que provienen, responder a algunas de ellas no es sencillo. Es necesario, en muchos casos, explorar los datos, el contexto y los usuarios de la información.

# El contexto de la vista minable



# El contexto de la vista minable

**¿Qué parte de los datos es pertinente analizar?**

**Vista minable**: una vista minable consiste en una vista en el sentido más clásico de base de datos: una tabla, por tanto, ha de recoger toda (y sólo) la información necesaria para realizar la tarea de minería de datos.

# El contexto de la vista minable

**¿Qué tipo de conocimiento se desea extraer y cómo se debe presentar?**

**Tarea, método y presentación**: Se trata de decidir qué tarea (clasificación, regresión, agrupamiento, reglas de asociación, etc.), cuáles son las entradas y las salidas (en las tareas predictivas), con qué método, entre los existentes para cada tarea (árboles de decisión, redes neuronales, regresión logística, etc.) y de qué manera se van a presentar o se van a navegar los resultados (gráficamente, como un árbol, como un conjunto de reglas, etc.)

# El contexto de la vista minable

**¿Qué conocimiento puede ser válido, novedoso e interesante?**

**Criterios de calidad**: En muchos casos hay que establecer unos criterios de comprensibilidad de los modelos (número de reglas máximo), criterios de fiabilidad (basados en medidas como la confianza para las reglas de asociación, la precisión para la clasificación, el error cuadrático medio para la regresión, etc.), criterios de utilidad (basados en medidas de cuándo son aplicables, como el soporte, qué beneficios se obtiene, a partir de matrices de costes, etc.), y criterios de novedad o interés (basados en medidas más o menos subjetivas).

# Reconocimiento

**Dominio y usuarios**

El procedimiento más similar a este reconocimiento es el establecimiento de requerimientos realizado por un analista funcional de software, aunque no buscaremos aquí casos de uso, sino que buscaremos los casos de usos y escenarios de las tomas de decisión.

- ¿Quiénes toman las decisiones?
- ¿Qué aspectos son cruciales en su negocio?
- ¿Existen decisiones que se toman de una manera arbitraria o basándose en reflexiones personales no explícitas?
- ¿Existe documentación sobre decisiones anteriores?
- ¿Qué decisiones son críticas? ¿Los modelos deben ser comprendidos y validados por expertos?
- ¿Qué conocimiento previo suele utilizar para apoyarse en sus decisiones? ¿Utiliza otras fuentes de datos externas para fundamentarse en sus decisiones?

# Reconocimiento

**Dominio y usuarios**

El resultado de este “reconocimiento” puede resumirse en una documentación u organizarse de una manera esquemática, estableciendo prioridades de análisis, destacando aquellas reglas de decisión importantes, que pueden mejorarse de manera significativa.

En general, se van descubriendo mayores posibilidades a medida que se va conociendo el dominio. Como hemos dicho al principio de este capítulo, sin este reconocimiento es imposible esclarecer las tareas, los métodos, los criterios de calidad, explorar los datos y el conocimiento previo.

# Exploración

**Exploración de los datos**

- Objetivo: Obtener una "vista minable" de los datos, lista para los modelos, no solo un resumen.
- Requisito Clave: Entender el negocio y el significado de los datos para saber qué buscar.
- Técnicas Esenciales:
  - Visualización: Usar gráficos (histogramas, box plots, etc.) para ver tendencias y anomalías.
  - Descripción: Resumir los datos con estadísticas (medias, máximos, nulos, etc.).
  - Agregación y Selección: Simplificar y elegir solo la información relevante para el análisis.

# Exploración

**Visualización**

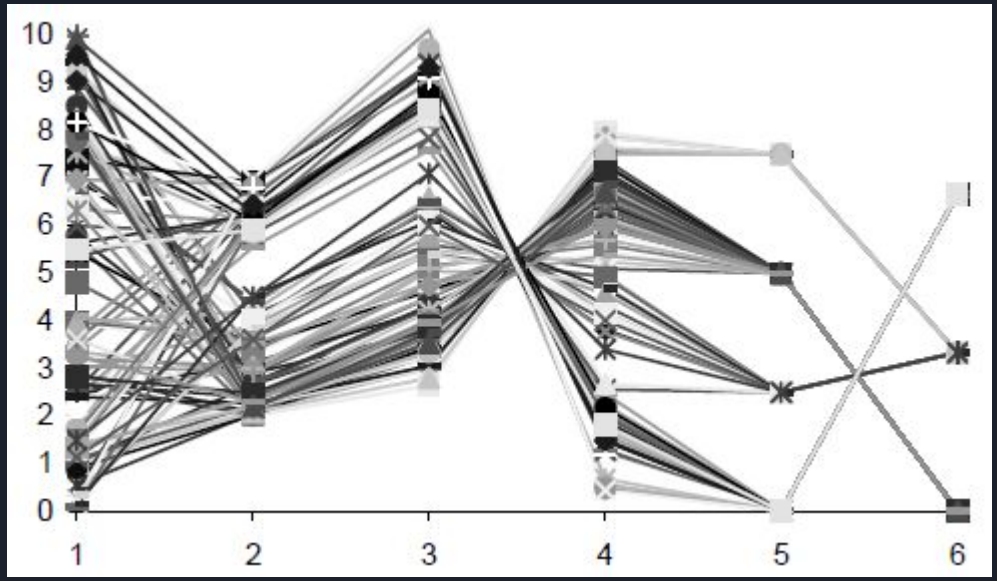
- **Objetivo principal**: Aprovechar la capacidad humana de ver patrones y anomalías.
- **Visualización Previa** (Minería de Datos Visual):
  - Explorar datos para encontrar tendencias y resúmenes.
  - Identificar posibles patrones para decidir qué método de minería usar.
- **Visualización Posterior**:
  - Validar los patrones descubiertos con expertos del dominio.
  - Facilitar la comprensión y comunicación de los resultados del modelo.

# Exploración

**Visualización multidimensional**

En la minería de datos, a menudo trabajamos con muchísimas variables, o lo que llamamos dimensiones. Visualizar tres dimensiones es fácil, pero ¿qué pasa cuando tenemos 4, 10, o 100? Es ahí donde entran las técnicas de visualización multidimensional. La más conocida y potente es la de coordenadas paralelas.

# Exploración



# Sumarización, descripción, generalización y pivotamiento

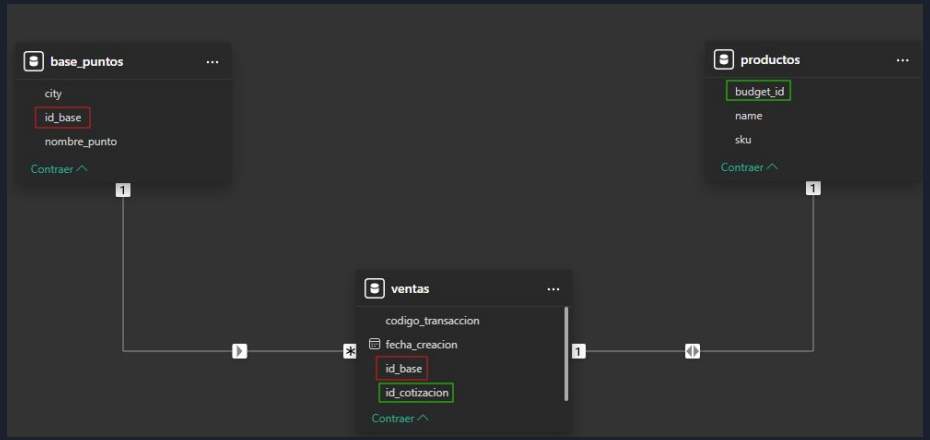
**Introducción**

**Problema**: Los datos históricos y multidimensionales no están en una única tabla, pero los modelos de minería de datos la necesitan.

**Solución**: Construir una "vista minable" unificando y resumiendo los datos.

**Operaciones esenciales**: JOIN (concatenar tablas), SELECT (elegir columnas), WHERE (filtrar filas) y GROUP BY (sumarizar y agrupar).

# Sumarización, descripción, generalización y pivotamiento




| codigo_transaccion | fecha_creacion | sku | nombre_punto | name | city |
|---|---|---|---|---|---|
| 734JEZ5J | martes, 01 de julio de 2025 | 134 | ALOT Staples - Barrio Norte II | Tarjetas innominadas | Recoleta |
| 734JEZLM | martes, 01 de julio de 2025 | 134 | Maxikiosco Madlen | Tarjetas innominadas | Ezeiza |
| 734JEZOM | martes, 01 de julio de 2025 | 134 | Cisplatina - Mario Bravo | Tarjetas innominadas | Almagro |
| 7GGVRO3Y | martes, 01 de julio de 2025 | 134 | Cangurito - Bahia Blanca - 32728 | Tarjetas innominadas | Bahía Blanca |
| 7GGVRO53 | martes, 01 de julio de 2025 | 134 | Triada Librería | Tarjetas innominadas | Villa Ballester |
| 7GGVROWY | martes, 01 de julio de 2025 | 134 | Encomiendas Tesei | Tarjetas innominadas | Villa Tesei |
| 7W3ZV25M | martes, 01 de julio de 2025 | 134 | MyC Maxikiosco | Tarjetas innominadas | Quilmes |
| 7W3ZVR64 | martes, 01 de julio de 2025 | 134 | ALOT Staples - Villa Crespo II | Tarjetas innominadas | Villa Crespo |
| 7W3ZVRB6 | martes, 01 de julio de 2025 | 134 | Computec I | Tarjetas innominadas | José C. Paz |
| 7W3ZVRE4 | martes, 01 de julio de 2025 | 134 | Punto Pacheco | Tarjetas innominadas | General Pacheco |
| 7W3ZVRK2 | martes, 01 de julio de 2025 | 134 | Servicios Del Interior - CENTRO | Tarjetas innominadas | Córdoba |
| 7W3ZVRMO | martes, 01 de julio de 2025 | 134 | Poochini Pets I | Tarjetas innominadas | Morón |

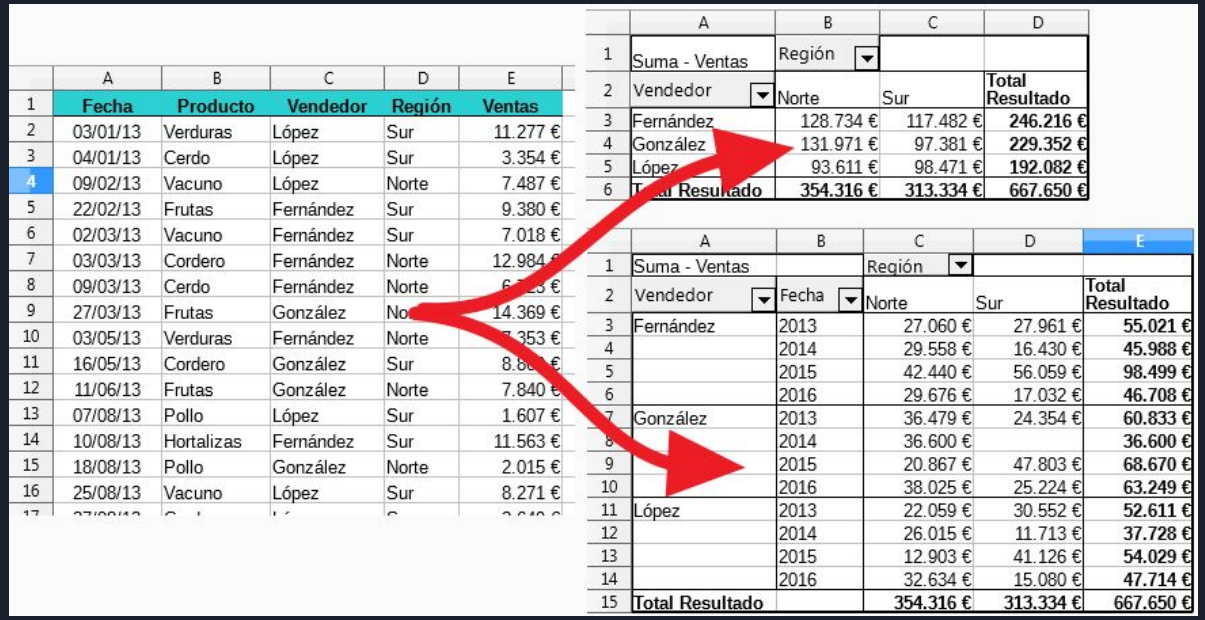
# Sumarización, descripción, generalización y pivotamiento

**Sumarización**

- ¿Qué es? Resumir datos detallados en valores agregados (promedio, total, etc.) para ver patrones.
- ¿Por qué es clave? Elimina el ruido de los datos. En datos muy detallados, los patrones se pierden; la agregación los hace visibles.
- Aplicación Práctica:
  - Crear nuevas variables: Generar atributos significativos (ej., "gasto total por cliente") a partir de datos crudos.
  - Análisis Exploratorio: Comparar grupos de datos para entender sus diferencias (ej., gasto promedio de clientes "activos" vs. "inactivos").

# Sumarización, descripción, generalización y pivotamiento

**Sumarización**



| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | Fecha | Producto | Vendedor | Región | Ventas |
| 2 | 03/01/13 | Verduras | López | Sur | 11.277 € |
| 3 | 04/01/13 | Cerdo | López | Sur | 3.354 € |
| 4 | 09/02/13 | Vacuno | López | Norte | 7.487 € |
| 5 | 22/02/13 | Frutas | Fernández | Sur | 9.380 € |
| 6 | 02/03/13 | Vacuno | Fernández | Sur | 7.018 € |
| 7 | 03/03/13 | Cordero | Fernández | Norte | 12.984 € |
| 8 | 09/03/13 | Cerdo | Fernández | Norte | 6.[illegible] € |
| 9 | 27/03/13 | Frutas | González | No[illegible] | 14.369 € |
| 10 | 03/05/13 | Verduras | Fernández | Norte | 7.353 € |
| 11 | 16/05/13 | Cordero | González | Sur | 8.8[illegible] € |
| 12 | 11/06/13 | Frutas | González | Norte | 7.840 € |
| 13 | 07/08/13 | Pollo | López | Sur | 1.607 € |
| 14 | 10/08/13 | Hortalizas | Fernández | Sur | 11.563 € |
| 15 | 18/08/13 | Pollo | González | Norte | 2.015 € |
| 16 | 25/08/13 | Vacuno | López | Sur | 8.271 € |

| | A | B | C | D |
|---|---|---|---|---|
| 1 | Suma - Ventas | Región | | |
| 2 | Vendedor | Norte | Sur | Total Resultado |
| 3 | Fernández | 128.734 € | 117.482 € | **246.216 €** |
| 4 | González | 131.971 € | 97.381 € | **229.352 €** |
| 5 | López | 93.611 € | 98.471 € | **192.082 €** |
| 6 | Total Resultado | **354.316 €** | **313.334 €** | **667.650 €** |

| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | Suma - Ventas | | Región | | |
| 2 | Vendedor | Fecha | Norte | Sur | Total Resultado |
| 3 | Fernández | 2013 | 27.060 € | 27.961 € | **55.021 €** |
| 4 | | 2014 | 29.558 € | 16.430 € | **45.988 €** |
| 5 | | 2015 | 42.440 € | 56.059 € | **98.499 €** |
| 6 | | 2016 | 29.676 € | 17.032 € | **46.708 €** |
| 7 | González | 2013 | 36.479 € | 24.354 € | **60.833 €** |
| 8 | | 2014 | 36.600 € | | **36.600 €** |
| 9 | | 2015 | 20.867 € | 47.803 € | **68.670 €** |
| 10 | | 2016 | 38.025 € | 25.224 € | **63.249 €** |
| 11 | López | 2013 | 22.059 € | 30.552 € | **52.611 €** |
| 12 | | 2014 | 26.015 € | 11.713 € | **37.728 €** |
| 13 | | 2015 | 12.903 € | 41.126 € | **54.029 €** |
| 14 | | 2016 | 32.634 € | 15.080 € | **47.714 €** |
| 15 | Total Resultado | | **354.316 €** | **313.334 €** | **667.650 €** |

# Sumarización, descripción, generalización y pivotamiento

**Generalización**

- Generalización: Simplificar los datos a un nivel más abstracto para encontrar patrones.
- Jerarquías de Valor: Pasar de un valor específico (9/7/2025) a un concepto (Día Festivo).
- Guía para el Analista: La generalización ayuda a decidir qué atributos agrupar, cuáles mantener y cuáles eliminar.
- Enriquecimiento: Proceso de añadir datos externos para crear jerarquías más significativas.

# Sumarización, descripción, generalización y pivotamiento

**Pivotamiento**

- ¿Qué es? Una operación que cambia filas por columnas para transformar radicalmente la estructura de los datos.
- Objetivo: Crear una vista minable única que los modelos de machine learning puedan usar.
- Ejemplo Clásico (Cesta de la Compra):
  - Antes: Una lista gigante con una fila por cada producto comprado.
  - Después: Una tabla única donde cada fila es una cesta y cada columna es un producto.

# Sumarización, descripción, generalización y pivotamiento

**Pivotamiento**

| Changuito | Prod_#1 | Prod_#2 | Prod_#3 | Prod_#4 | Prod_#5 |
|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 0 | 1 |
| 2 | 1 | 1 | 1 | 1 | 1 |
| 3 | 1 | 1 | 1 | 0 | 1 |
| 4 | 0 | 1 | 0 | 0 | 0 |
| 5 | 0 | 1 | 0 | 0 | 1 |
| 6 | 0 | 1 | 1 | 0 | 1 |
| 7 | 1 | 1 | 1 | 1 | 0 |
| 8 | 1 | 1 | 1 | 0 | 1 |
| 9 | 0 | 0 | 1 | 0 | 0 |
| 10 | 1 | 1 | 0 | 1 | 1 |

| Prod | Chango_#1 | Chango_#2 | Chango_#3 | Chango_#4 | Chango_#5 | Chango_#6 | Chango_#7 | Chango_#8 | Chango_#9 | Chango_#10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 |
| 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 |
| 3 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 |
| 4 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| 5 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 1 |

# Selección de Datos: Menos es Más 🎯

- ¿Qué es? Un paso crucial para elegir las filas y columnas correctas de tu tabla de datos.
- Objetivo:
  - Reducir el tamaño: Hacer los datos más manejables y los modelos más rápidos.
  - Mejorar el rendimiento: Eliminar el ruido y quedarte solo con la información más valiosa.

**Tipos de Selección:**

- Selección Horizontal (Muestreo):
  - ¿Qué hace? Reduce la cantidad de filas (ejemplos o individuos).
  - Ejemplo: Analizar solo el 10% de tus clientes en lugar del 100%.
- Selección Vertical (Reducción de Dimensionalidad):
  - ¿Qué hace? Reduce la cantidad de columnas (atributos o variables)
  - Ejemplo: Eliminar atributos irrelevantes como el número de teléfono o usar PCA para simplificar las variables.

# Selección de datos

**Técnicas de Muestreo: La Muestra para Ver el Todo 🎲**

- Objetivo: Reducir el tamaño de los datos para acelerar los modelos y, a menudo, mejorar la precisión.
- ¿Por qué? Una muestra bien elegida puede ser tan robusta como el conjunto completo y evita el ruido.

# Selección de datos

**Tipos de Muestreo**

- Aleatorio Simple:
  - ¿Qué es? Cada registro tiene la misma oportunidad de ser elegido.
  - Cuidado: Puede generar sesgo si los datos no son homogéneos.
- Aleatorio Estratificado:
  - ¿Qué es? Se divide la población en grupos (estratos) y se muestrea cada uno por separado.
  - Uso: Ideal para asegurar la representación de clases minoritarias (ej., casos de fraude).
- De Grupos:
  - ¿Qué es? Se seleccionan registros de grupos específicos y se descartan otros.
  - Uso: Para enfocarse en subpoblaciones relevantes para el análisis.