#2 Transformación de datos

Análisis y Exploración de datos | ITFS24

Prof. Martín Pasztetnik

¿Qué es la preparación de los datos y para qué sirve?

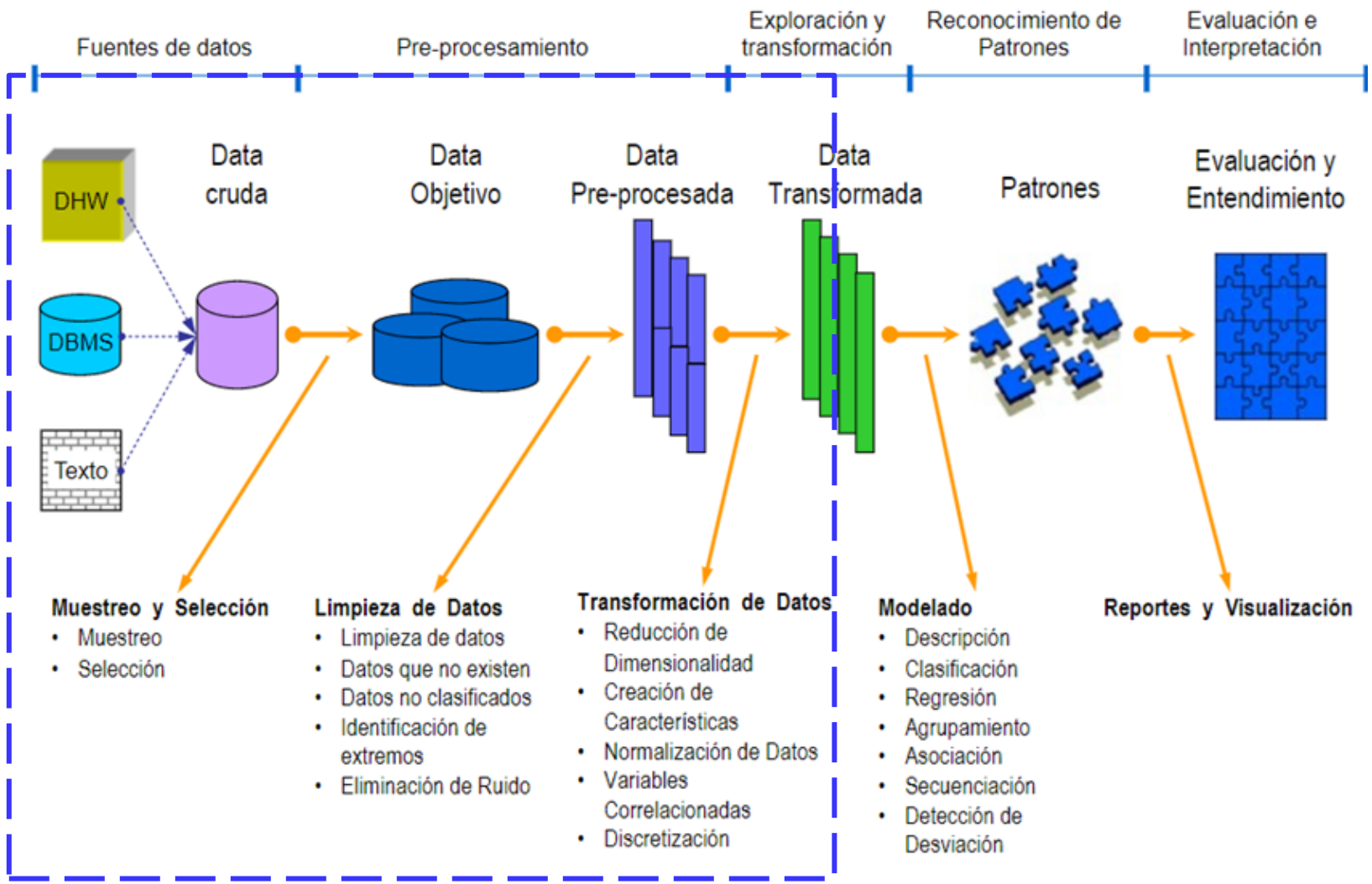
Durante la preparación de datos (también conocido como el “preprocesamiento”):

- consolidamos datos de distintas fuentes
- corregimos (limpiamos) problemas en nuestros datos que encontramos durante nuestra auditoría de los datos
- transformamos los datos al formato final que necesitamos para nuestro análisis.

Es una de las **tareas esenciales del análisis de datos.**

Objetivo: asegurar que los datos sean exactos y consistentes, hacerlos analizables, y no perder tiempo y recursos.

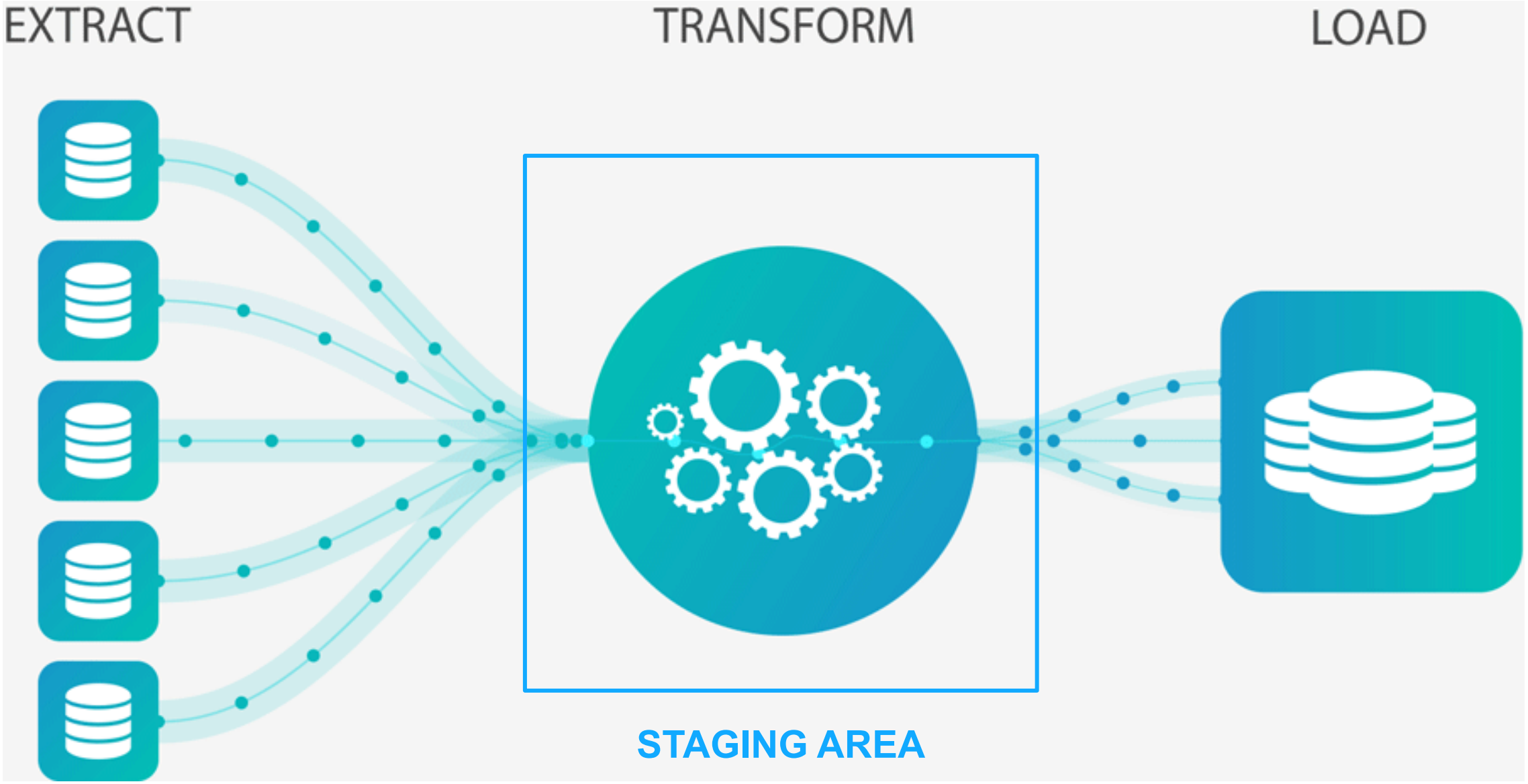
Ciclo de vida del Dato



Objetivo del ETL (Extract, Transform & Load)

ETL es un proceso de integración de datos que combina datos de múltiples fuentes de datos en un único almacén de datos coherente que se carga en un sistema de destino.

Igual que el preprocesamiento en general, el objetivo de ETL es producir **datos limpios y accesibles** que puedan utilizarse para hacer análisis, o para emplear en un sistema operacional, para apoyar un proceso de negocio.


EXTRACT
TRANSFORM
LOAD
STAGING AREA

Extracción

Los datos en bruto **deben extraerse de una variedad de fuentes,** por ejemplo:

- Bases de datos existentes
- Sistema CRM o ERP
- Registros de actividad aplicaciones, como el tráfico de red, informes de errores, etc.
- Sensor data
- Páginas web

Extracción parcial: solo extraer nuevos registros

Extracción completa: extraer todos los registros (y comparar con el último extracto para identificar los cambios que se han realizado)

Transformación

En el Staging Area, los datos **se transforman y consolidan para su caso de uso analítico previsto.**

Esta fase puede implicar las siguientes tareas:

- Filtrar, unir tablas, eliminar duplicados, validar y dar formato a los datos para que coincidan con el esquema del almacén de datos de destino.
- Realizar cálculos y traducciones basados en los datos sin procesar.
  Por ejemplo: cambiar los encabezados de filas y columnas para mantener la coherencia y convertir monedas u otras unidades de medida.
- **Realización de auditorías para garantizar la calidad de los datos**
- Eliminar, cifrar o proteger datos regidos por reguladores gubernamentales o de la industria.

Carga (Load)

La última fase de un proceso de ETL típico es **la carga de esos datos extraídos y transformados a su nuevo destino.**

Por lo general, esto implica una carga inicial de todos los datos, seguida de una carga periódica de cambios de datos incrementales y, con menos frecuencia, actualizaciones completas para borrar y reemplazar datos en el almacén.

La forma en que se cargan los datos puede diferir ampliamente.

- Algunos procesos pueden sobrescribir datos.
  Ejemplo: datos de contactación de clientes
- Otros procesos agregan nuevos datos en forma histórica.
  Ejemplo: un historial de venta

Limpieza de datos

El objetivo de la limpieza de datos es **eliminar ruido y resolver las inconsistencias**.

En este proceso se desestiman los datos erróneos, atributos que no suman al análisis, y se validan aquellos que son útiles.

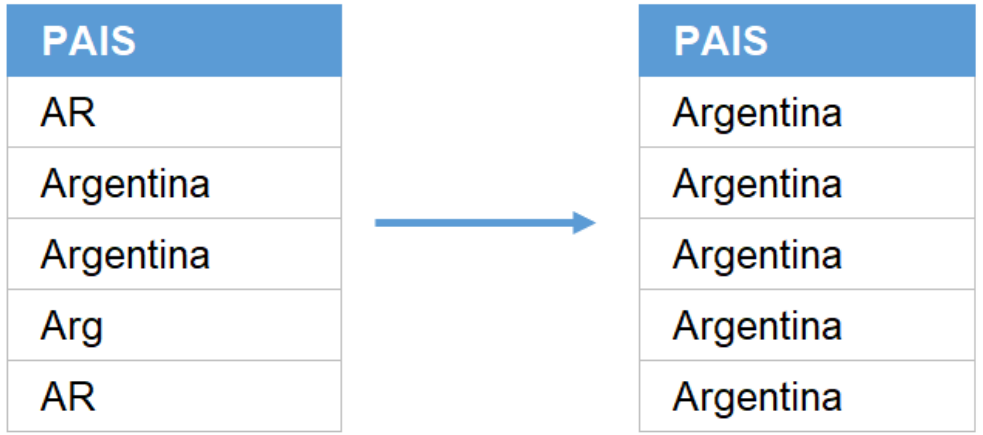
Limpieza: normalización de datos

Se identifican aquellos valores que, siendo iguales, aparecen con notaciones o nomenclaturas diferentes y se los **reescribe de una manera uniforme**. También se identifican las variables categóricas y se corroboran que sean uniformes

Ej: CALLE S MARTIN, CALLE GRAL. SAN MARTÍN, CALLE JOSE DE SAN MARTIN...
= CALLE GRAL JOSE DE SAN MARTIN

Ej 2: Femenino/Masculino/M/F/Mujer/...

Ej 3: Arg, AR y Argentina.



Limpieza: correcciones de formato

El objetivo es **resolver problemas de formato y asignar los tipos correctos de datos**, y corregirlos para hacerlos compatibles con la estructura de las bases de datos: tipología, formatos y codificación.

¿A que refiere un **tipo de dato**? Por ejemplo:

- String/Character: “hola”, “Argentina”
- integer: 7, 22, 1078
- float: 1.5, 16.788
- boolean: TRUE, FALSE
- date: 2022-07-01, 7/1/2022

¿Por qué es importante?

El formato en que se encuentran los datos va a afectar nuestro análisis por varias razones. Por ejemplo, las operaciones que se pueden realizar dependen del tipo de datos. Además algunos tipos ocupan menos espacio en memoria que otros.

Ej 1: Formatos de fechas 2003-09-01 en vez de 200309010000.
Ej 2: Números interpretados como texto
Ej 3: Formatos (UTF8)

Limpieza: imputación de valores perdidos

Los datasets siempre suelen venir con datos faltantes que responden a información que se perdió o nunca se recolectó. Existen varias técnicas para **completar datos faltantes**. Al proceso de completar datos faltantes se lo llama “imputación”.

- Debemos poder detectar, rellenar o eliminar datos faltantes.
- Hay que tener en cuenta porque faltan los datos (es al azar o no?)
- Hay que utilizar conocimiento del dominio para definir cuáles datos faltantes se completarán y cómo.

Posibles formas de imputación:

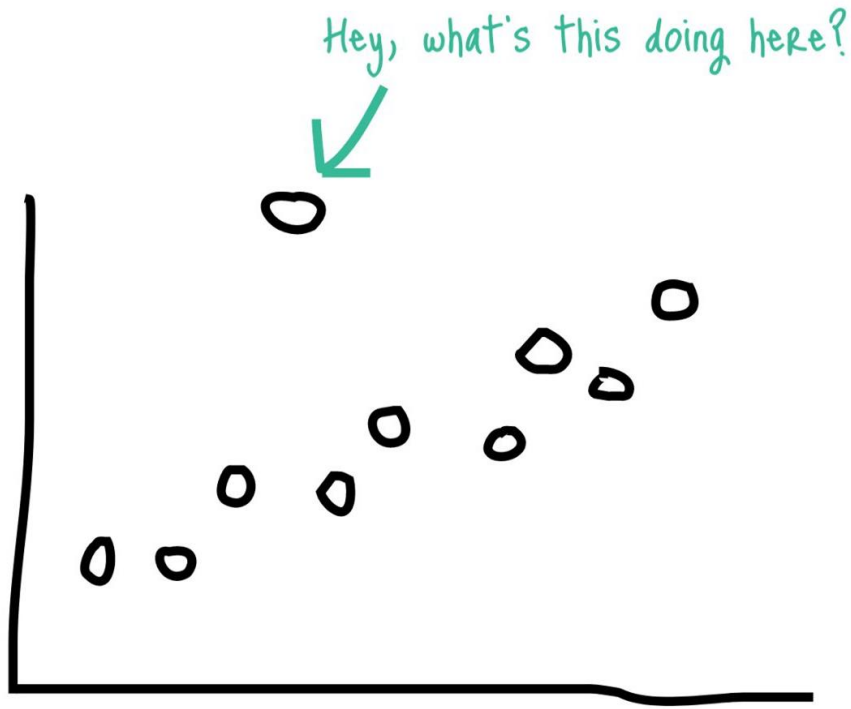
- Valor constante (0, “Desconocido”)
- Media, mediana o valor más frecuente
- “Last value carried forward” (LOCF)
- Usando la predicción de una regresión u otro modelo
- y hay varios métodos más avanzados…

Limpieza: identificación de ruido

Identificamos resultados inesperados, por ejemplo números mucho más altos o bajos de lo esperado. Cuando un valor se aparta notoriamente del comportamiento general, lo llamamos un **outlier (valor atípico).**

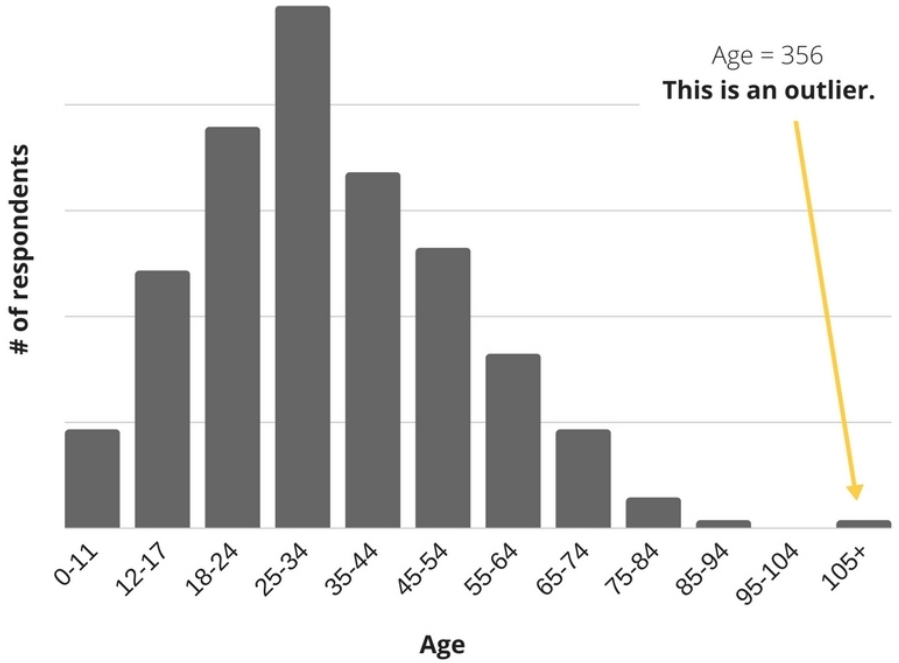
Un outlier puede ser el resultado de un error de los datos o puede ser un dato correcto que representa una anomalía en la realidad.

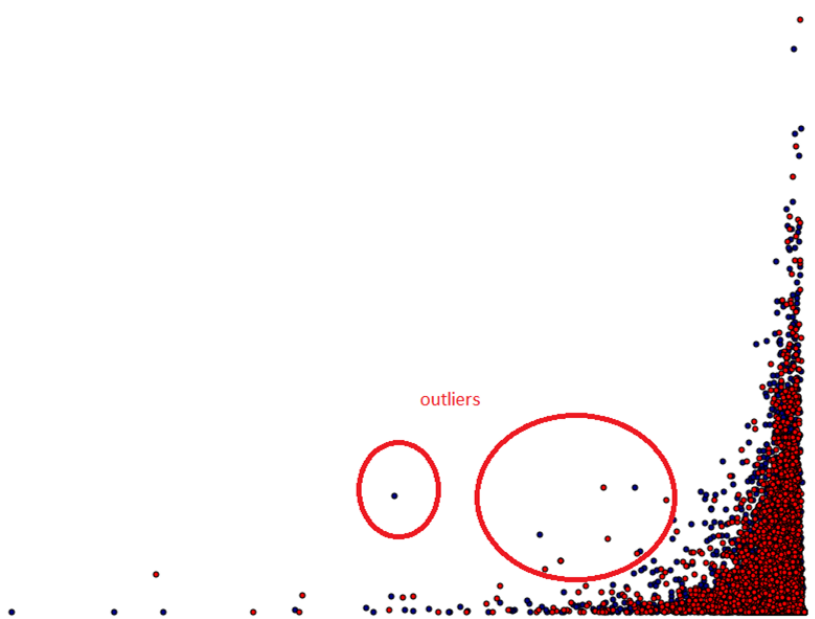
Ej: corrección de errores como un campo de edad con valores negativos o mayores a 100

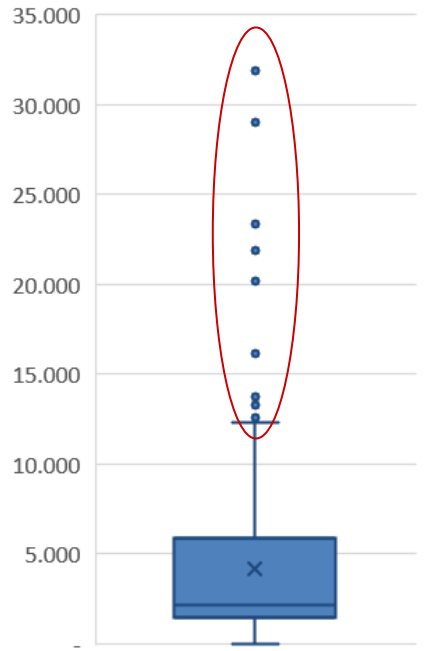


Limpieza: identificación de ruido

Para identificar valores atípicos se suelen usar visualizaciones como un histograma, un scatter plot o un box plot. También se puede usar métodos estadísticos para identificar los outliers, por ejemplo el rango intercuartílico o z-scores.







¿Qué es la reducción de datos?

Muchas veces obtendremos datos que son abarcativos, con muchos registros y muchas variables/columnas. **Deberemos aprender a elegir, segmentar, seleccionar lo que nos interesa**.

La **reducción**, es una forma de preparación y transformación de los datos, que busca **minimizar la dimensión** de los datasets, reduciendo su tamaño o sus variables a partir de diversas técnicas. Esta reducción **no** es aleatoria y responde a las necesidades analíticas.

Reducción de datos

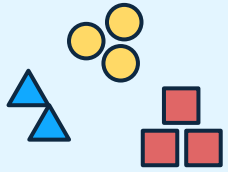
Objetivo: minimizar el número de variables y/o registros para el análisis, filtrando y combinando variables únicas en variables compuestas.

Algunas formas de reducir los datos:

FILTRAR

SELECCIONAR

AGRUPAR/ Sumarizar



Recordemos el “ABC” de las tablas.

Cada **fila** es un registro/una observación

Cada **columna** es una variable/un atributo

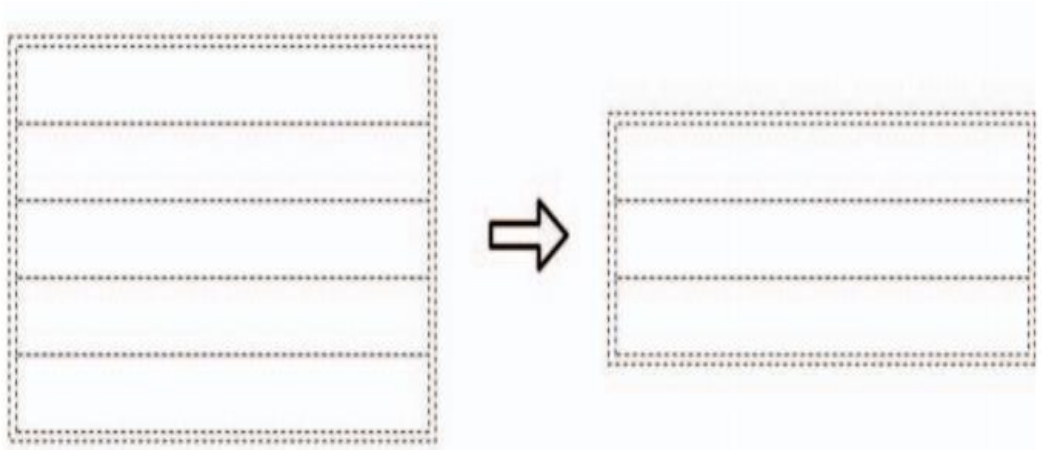
La intersección entre una fila y una columna, es un dato y es la categoría/magnitud que adquiere esa variable

Cada **tabla** es una entidad

Filtrado de registros

¿Qué es?
Cuando filtramos registros, estamos eliminando filas de nuestro dataset.

Objetivo:
Reducir el universo de datos a analizar. Esto puede ser útil por ejemplo porque parte de los datos no son de interés para nuestro análisis, o porque el dataset es muy pesado y el tamaño complicaría nuestro análisis.

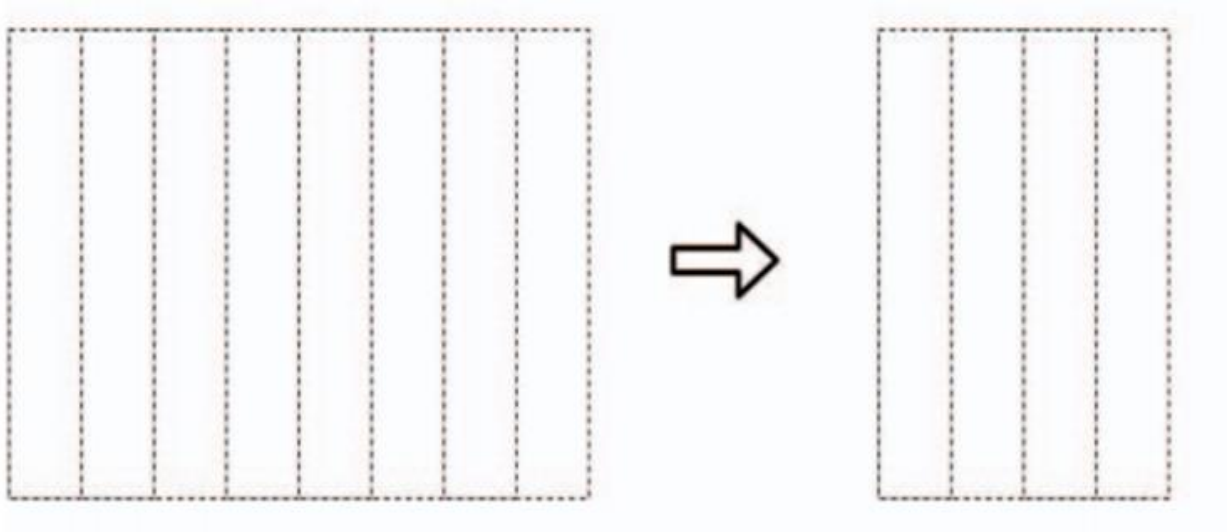


Ejemplo: imaginemos que trabajamos en Adidas, y queremos evaluar la venta de zapatillas deportivas: sobre una tabla de productos, solo me interesarían los productos de un determinado rubro, así que los filtro por esa categoría.

Selección de atributos/variables/columnas

¿Qué es?
Cuando seleccionamos atributos, estamos eliminando columnas de nuestro dataset.

Objetivo:
Reducir la cantidad de características de nuestro universo a analizar. El objetivo es quedarnos solo con los atributos que aportan a nuestro análisis



Ejemplo: dada una necesidad de segmentación, sobre una tabla de clientes, solo me interesan sus atributos de segmento: edad, sexo y domicilio.

Agrupar datos

Existen muchas técnicas distintas para agrupar datos y los objetivos por lo cual agrupar los datos pueden ser distintos.

Algunas técnicas **agrupan registros**, por ejemplo para poder sumarizar datos y permitir comparaciones entre ciertos grupos.

Otras técnicas **agrupan atributos**, por ejemplo combinar atributos que nos dan información similar para bajar la cantidad de atributos en nuestro dataset.

Agrupar datos: discretización

Discretización es la tarea de reducir el número de valores de una variable continua agrupándolos en una serie de intervalos (bins).

Cuando discretizamos una variable perdemos cierto detalle de información, pero nos puede ayudar a sumarizar información, comparar ciertos grupos de registros y visualizar información en una forma más amigable.

Peso en Kilogramos	Bins de Peso
60,3	60-70 kgs
70,8	70-80 kgs
65,6	60-70 kgs
92,4	90-100 kgs
58,5	50-60 kgs

Agrupar datos: clustering

Clustering es la tarea de agrupar datos por variables similares. La idea es que los registros en el mismo grupo (llamado clúster) son más similares entre sí que a los de otros grupos.

En el mundo empresarial se usa clustering por ejemplo para crear distintos segmentos (perfiles) de clientes para campañas de marketing.