Preprocesamiento y estructuración de datos I

Tema 3.

unir LA UNIVERSIDAD EN INTERNET

Profesor: Luis F. Ruiz

¿Qué veremos hoy?

01 Intro y objetivos

02 Principios de organización de datos

03 Limpieza y transformación en herramientas visuales

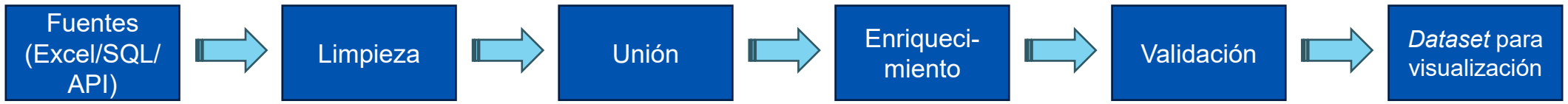
04 Procesos de preprocesamiento con lenguajes de programación

05 Herramientas

3.1. Introducción y Objetivos

- ✓ ¿Por qué el preprocesamiento condiciona la visualización? (***Regla del 70%)***
- ✓ Impacto en calidad, rendimiento y confianza del usuario
- ✓ Objetivo: flujo reproducible, trazable y eficiente

Esquema general de preprocesamiento



3.2. Principios de organización de datos

- ✓ Estructura (***tidy data***) ➔ Reglas fundamentales (Hadley Wickham*):
  - ➢ una **fila=observación**,
  - ➢ una **columna=variable**
  - ➢ Claves/relaciones claras entre tablas (ID)
- ✓ Convenciones: nombres consistentes, tipos correctos, categorías normalizadas

*Su trabajo está referenciado en la bibliografía junto con la obra:
"*R for Data Science: Import, Tidy, Transform, Visualize, and Model Data".*

3.2. Reglas prácticas (antes / después)

- ✓ Eliminar nulos críticos, resolver duplicados
- ✓ Normalizar categorías (espacios, mayúsculas, acentos)
- ✓ Documentar supuestos y decisiones de transformación

Ejemplo: Transformando a Tidy data (Pivote)

Ej. Ventas de 3 meses (Excel)

Producto	Enero	Febrero	Marzo
A	150	180	165
B	90	110	130

untidy data ➔ ***tidy data***

Desordenado ➔ Ordenado

Producto	Mes	Ventas
A	Enero	150
A	Febrero	180
A	Marzo	165
B	Enero	90
B	Febrero	110
B	Marzo	130

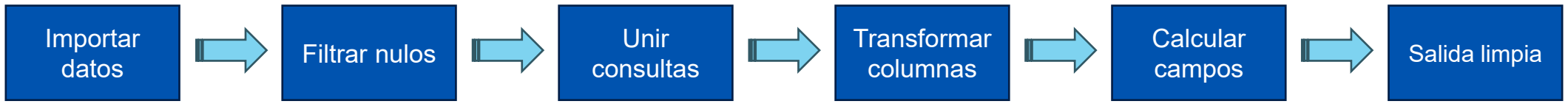
3.3. Limpieza y transformación en herramientas visuales

Herramienta	Entorno	Enfoque Principal
Power Query (M)	Integrado en Power BI	Procesamiento optimizado, trazabilidad visual y conexión amplia de fuentes.
Tableau Prep	Plataforma de Tableau	Flujos visuales interactivos (nodos), previsualización en tiempo real, automatización.
OpenRefine	Código Abierto	Limpieza masiva, estandarización y normalización avanzada (ej. clustering).

3.3. Limpieza y transformación en Power Query (Power BI)

- Importar ➔ Filtrar nulos ➔ Unir consultas ➔ Transformar columnas ➔ Calcular campos ➔ Salida
- Ventajas: M reproducible, UI guiada, integración con el modelo

Flujo de preprocesamiento en Power Query



3.3. Ejemplo Power BI (equivalencias con Python)

- Cambio de tipo, filtrado, columnas personalizadas, merges

Preprocesamiento en Python (pandas)

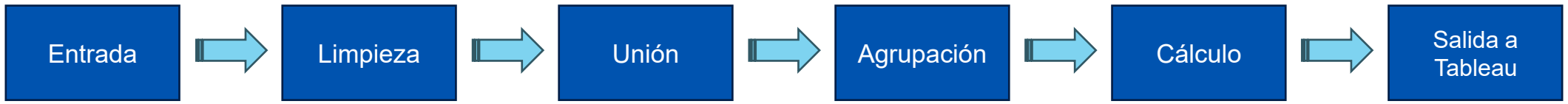
```
import pandas as pd

df = pd.read_excel("ventas.xlsx")
df = df.dropna(subset=["Importe"])
df["Margen"] = df["Ingreso"] - df["Coste"]
df_clean = df[df["Margen"] > 0]
df_clean["Canal"] = df_clean["Canal"].str.strip().str.title()
```

3.3. Limpieza y combinación en Tableau Prep

- Entrada ➔ Limpieza ➔ Unión ➔ Agrupación ➔ Cálculo ➔ Salida a Tableau
- Ventajas: flujo visual, perfiles de datos, pasos documentados

❖ **Enfoque Visual:** modelo basado en **flujos interactivos** (cada paso = un **nodo**).
❖ **Operaciones Comunes:** Unir tablas, agregar registros, **pivotado** (para crear *tidy data*) y crear campos calculados.
❖ **Integración y Automatización:**
  - Flujos de datos = extractos reutilizables en **Tableau Desktop**.
  - **Tableau Prep Conductor** = automatizar ejecución periódica de limpieza = actualización continua.

Flujo de limpieza y combinación en Tableau



3.3. Ejemplo Tableau Prep (equivalencias con R)

- Reglas claras y validación con perfiles de datos
- Buenas prácticas: nombres consistentes, categorías limpias, campos derivados

Preprocesamiento en R (dplyr)

```
library(readxl)
library(dplyr)
df <- read_excel("ventas.xlsx") %>%
  filter(!is.na(Importe)) %>%
  mutate(Margen = Ingreso - Coste,
         Canal = stringr::str_to_title(stringr::str_trim(Canal))) %>%
  filter(Margen > 0)
```

3.4. Procesos de preprocesamiento con lenguajes de programación

Lenguaje	Librería Clave	Funciones Típicas
R	Tidyverse (especialmente dplyr)	filter(), mutate() (crear variables), summarise() (agregar). Utiliza el operador de tubería %>% para encadenar pasos.
Python	Pandas (usa DataFrames)	Eliminación de duplicados, imputación de valores faltantes, cambio de tipos de datos, normalización de variables. Integración con librerías numéricas (NumPy) y de bases de datos.

3.4. Preprocesamiento con Python (pandas)

- Lectura, limpieza de nulos, cálculo de métricas, normalización de categorías
- Cuándo preferir Python: volumen, automatización, lógica compleja

Preprocesamiento en Python (pandas)

```
import pandas as pd

df = pd.read_excel("ventas.xlsx")
df = df.dropna(subset=["Importe"])
df["Margen"] = df["Ingreso"] - df["Coste"]
df_clean = df[df["Margen"] > 0]
df_clean["Canal"] = df_clean["Canal"].str.strip().str.title()
```

3.4. Preprocesamiento con R (dplyr)

- Pipelines declarativos, transformación legible y reproducible
- Cuándo preferir R: análisis estadístico, reproducibilidad, CRAN

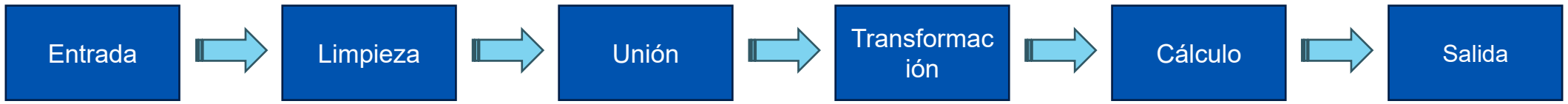
Preprocesamiento en R (dplyr)

```
library(readxl)
library(dplyr)
df <- read_excel("ventas.xlsx") %>%
  filter(!is.na(Importe)) %>%
  mutate(Margen = Ingreso - Coste,
         Canal = stringr::str_to_title(stringr::str_trim(Canal))) %>%
  filter(Margen > 0)
```

Comparativa: Power Query vs Tableau Prep vs Código

- **Power Query**: integración BI, reproducible, auditable
- **Tableau Prep**: claridad de flujo, perfiles de datos
- **Python/R**: flexibilidad, escalabilidad y automatización

Comparativa de flujos: Power Query vs Tableau Prep



Un paso más en validación y control de calidad de datos (no examen)

- **Práctica:**
  - Definir reglas explícitas para comprobar:
    - Presencia de valores nulos en campos clave.
    - Unicidad de las claves primarias.
    - Coherencia de rangos numéricos (ej., que las ventas no sean negativas).
  - **Documentación:** Registrar los criterios aplicados y los resultados de cada validación. Esto incrementa la fiabilidad de las visualizaciones.

- **Herramientas Específicas:**
  - En Python: la librería **pandera** permite definir esquemas de validación sobre los *DataFrames*.
  - En R: paquetes como **assertr** o **validate** facilitan la creación y aplicación sistemática de reglas.

Conclusiones

- ✓ Buen preprocesamiento = visualización fiable
- ✓ Define reglas, documenta pasos y valida con datos de prueba
- ✓ **Práctica**: replicar flujos en Power BI y Tableau Prep

Muchas gracias por vuestra atención

www.unir.net