Módulo 3

Sesión N° 6

ACTIVIDAD:

Integración y Auditoría de Datos en Pandas

Objetivo: Desarrollar un pipeline de integración, validación y reporte usando pandas, aplicando merge, groupby, pivot y melt. Demostrarás su capacidad para auditar la calidad de la integración y documentar el proceso.

Contexto

Trabajas como ingeniero/a de datos para una empresa nacional de retail que está implementando un nuevo sistema de inteligencia de negocio. Recibes archivos con ventas diarias, información de productos y detalles de tiendas. Tu desafío es integrar estos datos, validar la consistencia, detectar registros problemáticos y generar un reporte que pueda ser usado para la toma de decisiones ejecutivas.

Entregable:

- Un notebook o script con el código y comentarios.
- Un diagrama de flujo del pipeline (puede ser digital o foto de un esquema a mano).
- Un informe breve (1 – 2 páginas) justificando tus decisiones y mejoras.
- Formato: grupal.

Requerimientos:

1. Carga de datos
    - Simula la carga de tres DataFrames:
        - ventas (columnas: id_venta, id_producto, id_tienda, fecha, monto)
        - productos (columnas: id_producto, categoria, nombre)
        - tiendas (columnas: id_tienda, region, tipo_tienda)
    - Crea al menos 10 registros en cada tabla, asegurando que existan algunos id_producto y id_tienda sin correspondencia.
2. Integración y validación
    - Realiza los merges necesarios para obtener una tabla final que contenga todas las ventas con su información de producto y tienda.
    - Valida la unicidad de claves en cada tabla antes del merge.
    - Detecta y reporta:
        - Ventas sin producto asociado
        - Ventas sin tienda asociada
        - Duplicados en claves
    - Genera un DataFrame aparte con los registros “huérfanos” y explica posibles causas.
3. Agrupamiento y reporte
    - Agrupa las ventas por region, categoria y año (extraído de la columna fecha), calculando:
        - Suma total de ventas
        - Monto promedio por transacción
        - Número de ventas
    - Genera una tabla dinámica (pivot) con región como filas, años como columnas y ventas totales como valores.
    - Despivotea la tabla para preparar un reporte largo listo para modelado.