Módulo 3

Sesión N° 3

ACTIVIDAD:

Integración y Automatización de Extracción de Datos desde Archivos CSV, Excel y Web

Objetivo: Aplicar técnicas de extracción, integración y validación de datos provenientes de archivos CSV, Excel y web, automatizando el flujo con Python y Pandas.

Contexto

Automatiza el proceso de extracción de datos desde un conjunto de archivos CSV y Excel ubicados en un directorio, integra la información en un solo DataFrame, y exporta los resultados consolidados en un archivo Excel, incluyendo una hoja resumen con estadísticas descriptivas y una hoja adicional con la extracción de una tabla relevante desde la web.

Requerimientos técnicos
- Python 3.x
- Pandas
- openpyxl
- Acceso a internet para la extracción web

Requerimientos:

1. Recorre todos los archivos CSV de un directorio y concatena su contenido en un solo DataFrame.
2. Recorre todos los archivos Excel (.xls, .xlsx) del mismo directorio, leyendo la hoja principal de cada uno y concatenando los datos en otro DataFrame.
3. Integra ambos DataFrames (CSV y Excel) en un solo DataFrame global.
4. Extrae una tabla de la web (por ejemplo, la tabla de países por población de Wikipedia) usando pd.read_html(). Agrega esta tabla como una hoja adicional en el archivo Excel final.
5. Genera una hoja resumen con estadísticas descriptivas básicas (describe()) del DataFrame global.
6. Exporta los resultados a un archivo Excel llamado consolidado_resultados.xlsx con al menos las siguientes hojas:
    - "Datos_CSV"
    - "Datos_Excel"
    - "Resumen"
    - "Web"
7. Comenta tu código, documenta supuestos y problemas encontrados.

Instrucciones de Desarrollo:
Modalidad grupal.
Tiempo: 100 minutos.
Entrega: archivo consolidado_resultados.xlsx con las hojas requeridas, y el script/documentación utilizados.