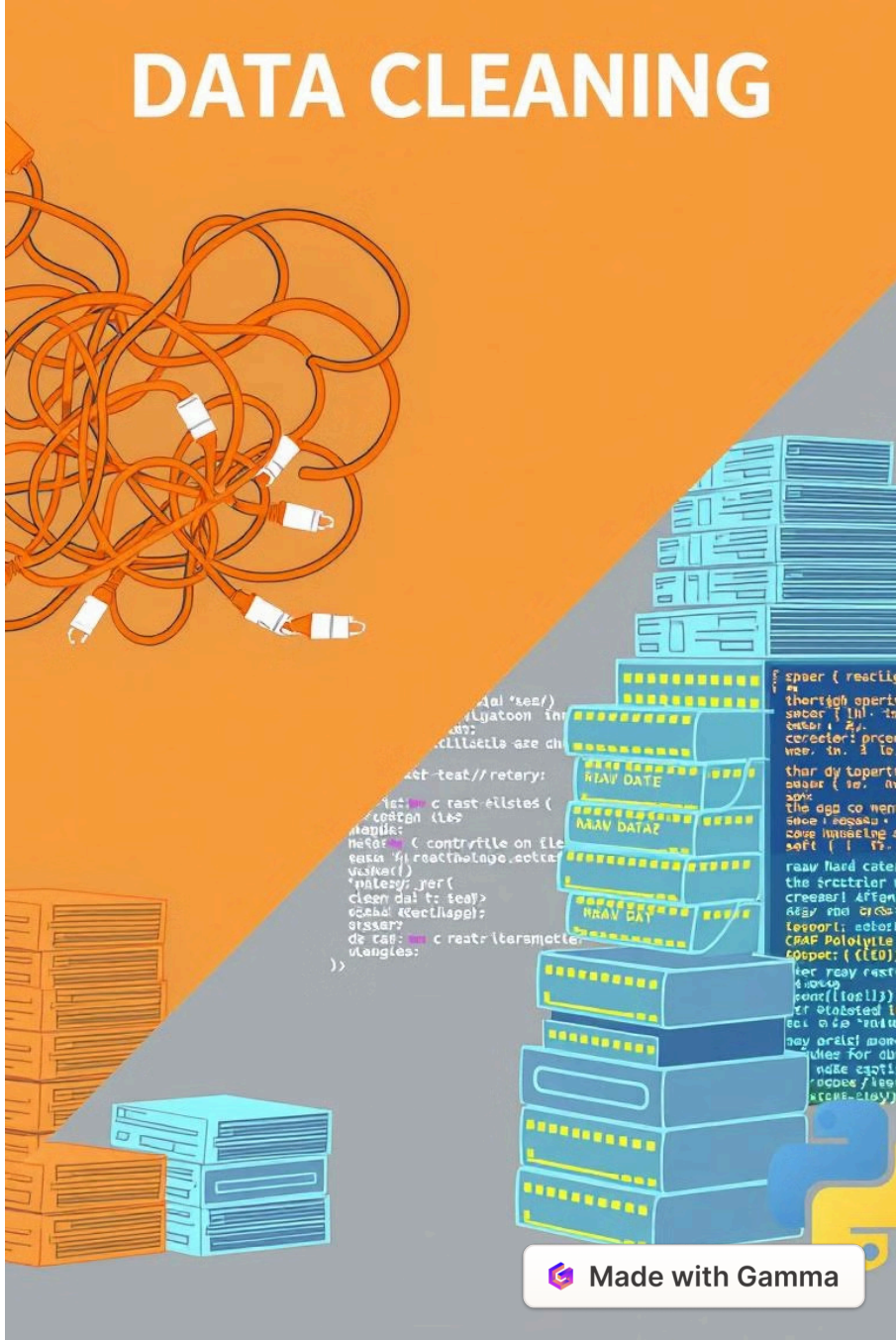
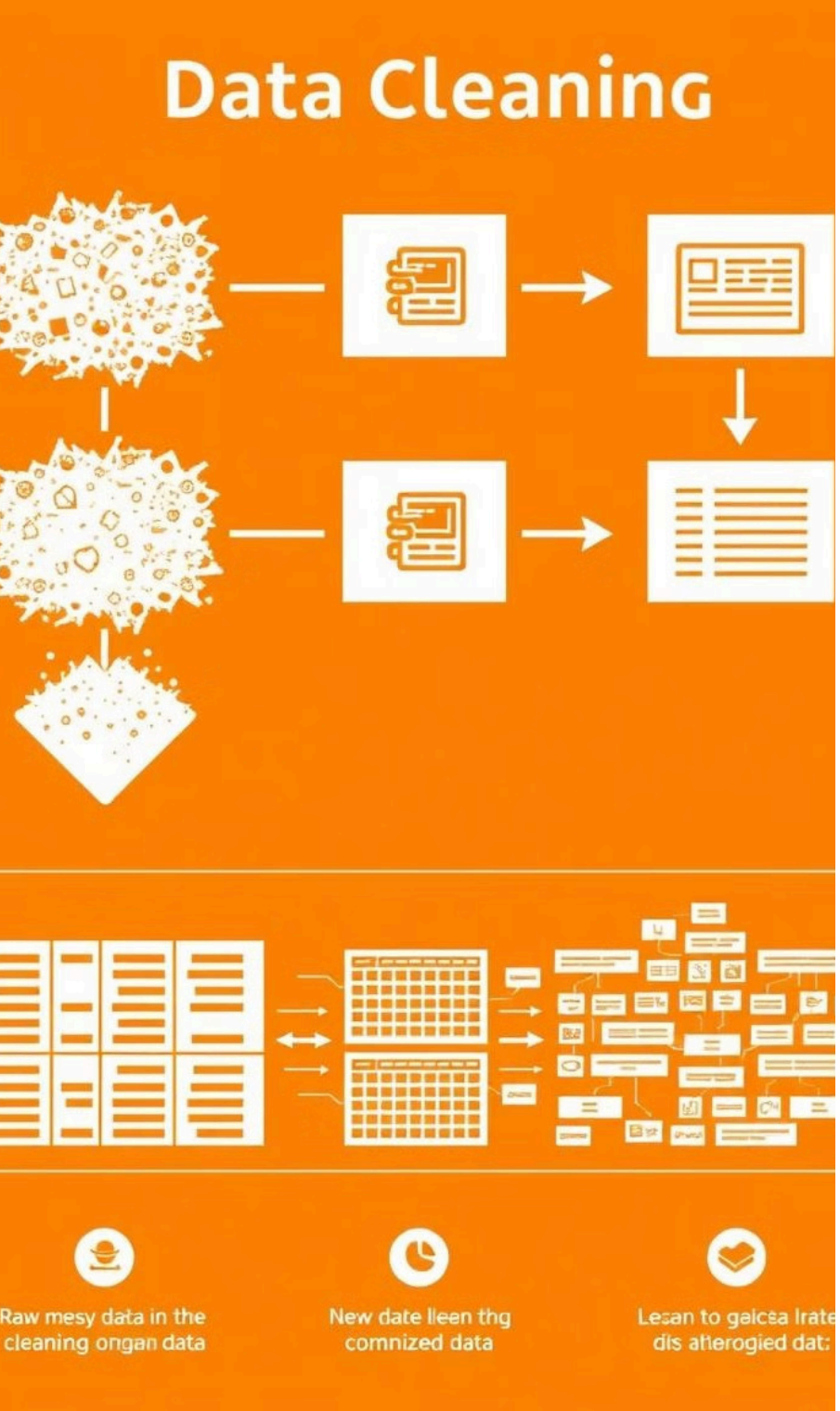
Data Cleaning with Python

Autores: Juan Carlos Macías y Fernando García Catalán





Definición e Importancia de la Limpieza de Datos

¿Qué es?

La limpieza de datos (Data Cleaning o Data Cleansing) es el proceso de detectar, eliminar, corregir o transformar cualquier anomalía, perturbación o irrelevancia de los datos.

¿Por qué es importante?

Es necesario antes de cualquier tarea de análisis, desde visualizaciones hasta modelos de machine learning. Ayuda a obtener resultados confiables, reduce posibles sesgos y es fundamental porque los datos en el mundo real rara vez son homogéneos y directamente intuitivos.

Consecuencias de no limpiar

Datos sin limpiar pueden llevar a conclusiones erróneas, tendencias falsas y estadísticas inexactas.

Tipos Comunes de Datos "Sucios"

Celdas vacías (Empty cells)

Celdas sin valor.

Datos en formato incorrecto

Datos que no cumplen el formato esperado. Ejemplo: una fecha como texto en lugar de formato fecha.

Datos erróneos

Datos que son incorrectos o fuera de rango según el contexto.

Datos duplicados

Filas que se han registrado más de una vez.

Valores nulos

Representaciones específicas de la ausencia de datos.

Valores atípicos (Outliers)

Valores que se desvían significativamente de la norma.

Incoherencias

Datos representados en diferentes unidades, formatos o estilos.

Limpieza de Datos con Pandas - Técnicas Específicas

Importación de Datos

Pandas se utiliza para leer archivos de datos en diferentes formatos (ej., pd.read_csv()) y convertirlos en DataFrames.

Detección de Valores Nulos

isnull(): Detecta valores nulos y devuelve un DataFrame booleano.

notnull(): Devuelve el resultado lógico contrario a isnull().

Tratamiento de Valores Nulos

dropna(): Elimina las filas que contienen valores nulos.

fillna(): Rellena los valores nulos con un valor predeterminado.

Imputación con la media, mediana o moda.

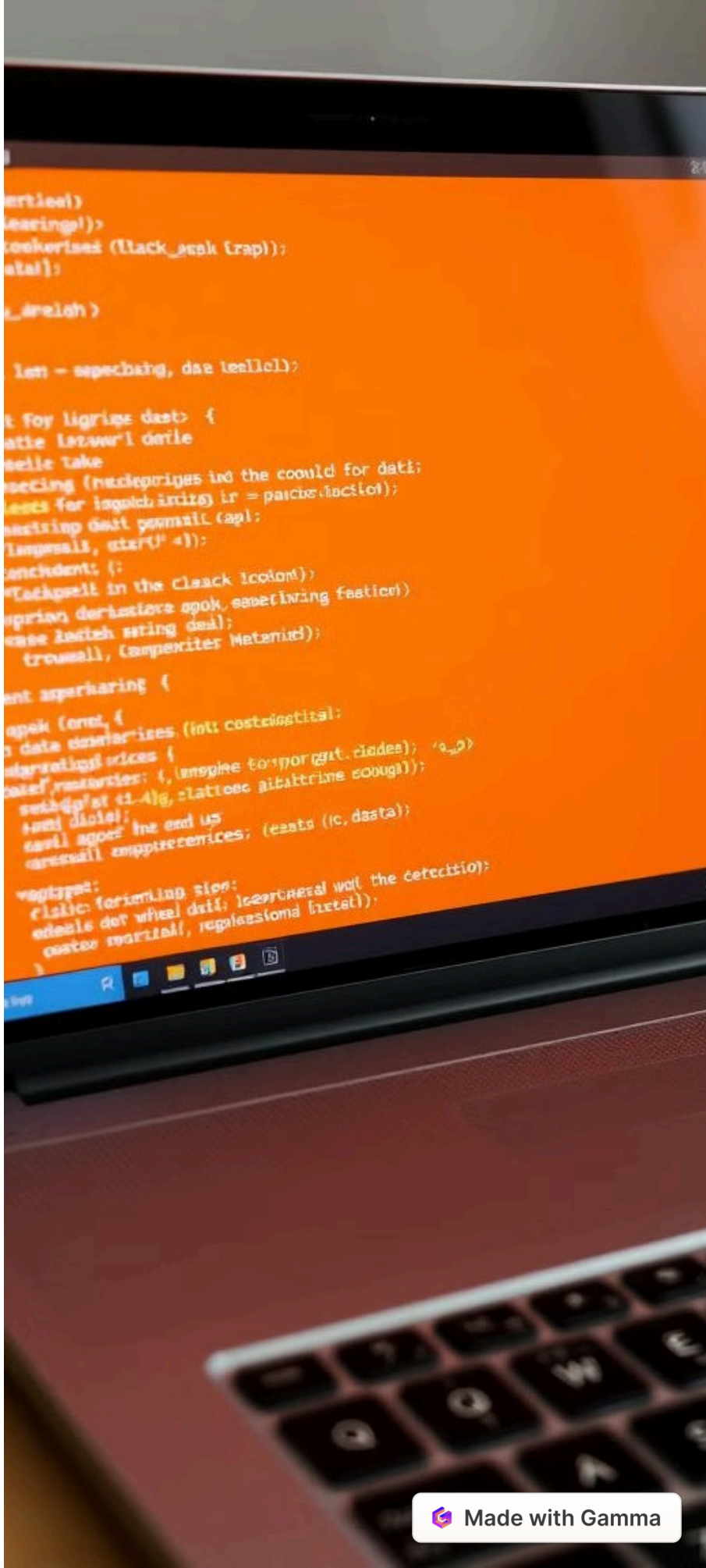
Imputación KNN o por regresión para métodos más complejos.

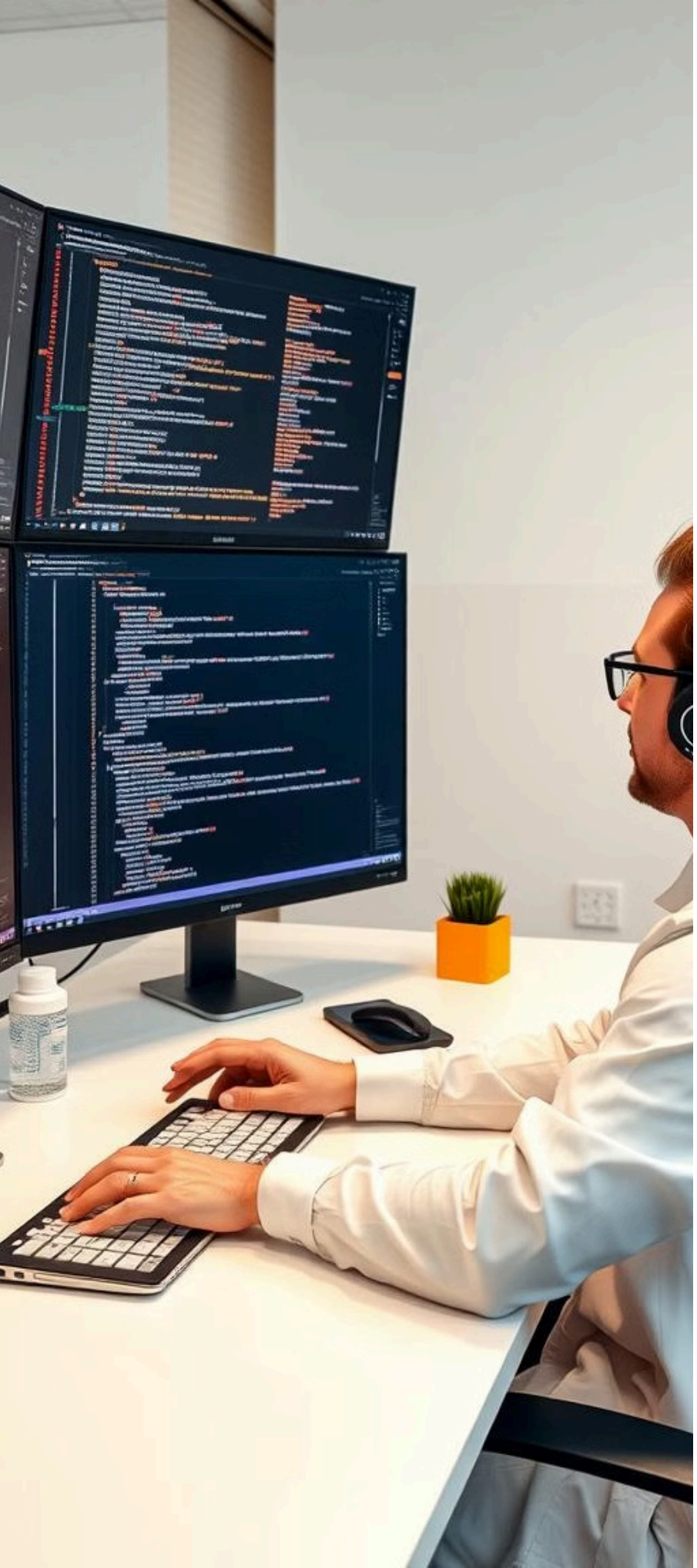
H6

Tratamiento de Datos en Formato Incorrecto

Conversión de tipos con pd.to_datetime().

Eliminación de filas con formato incorrecto.





Más Técnicas de Limpieza con Pandas

Tratamiento de Datos Erróneos

Identificación a través de inspección visual (.head()) o métodos descriptivos (describe()).

Reemplazo directo de valores incorrectos.

Eliminación de filas con datos significativamente erróneos.

Tratamiento de Duplicados

duplicated(): Devuelve una Serie booleana indicando qué filas son duplicados.

drop_duplicates(): Elimina las filas duplicadas.

Detección y Tratamiento de Valores Atípicos

Inspección visual con histogramas y boxplots.

Métodos estadísticos como Z-score e IQR. (bigotes)

Eliminación o transformación de outliers.

Manejo de Incoherencias

Normalización de formato (.strip(), .upper(), .lower(), .to_datetime(), .astype()).

Conversión de unidades y normalización de valores categóricos.

Técnicas Avanzadas y Operaciones con DataFrames

Eliminación de Columnas o Filas Irrelevantes

drop(): Eliminar columnas o filas innecesarias.

set_index(): Establecer una columna como índice para mejorar la eficiencia.

Operaciones con Strings

Métodos .str para manipulación de texto.

Reemplazo Condicional

Uso de np.where() para modificar valores basados en condiciones.

Aplicación de Funciones a DataFrames

applymap(): Aplica una función a cada elemento del DataFrame.

apply(): Aplica una función a lo largo de un eje del DataFrame.

Renombrar Columnas

Método rename() para hacer los nombres de las columnas más comprensibles.

Exploración de Datos y Feature Engineering

Exploración de Datos (EDA)

Antes de la limpieza, es crucial explorar los datos para comprender su estructura, identificar posibles problemas y obtener información valiosa. Esto incluye examinar la fuente de los datos en busca de posibles sesgos, entender el contexto, determinar el número de variables y categorías, observar estadísticas descriptivas y visualizar los datos. La EDA puede ser la primera indicación de datos sucios.

Ingeniería de Características

Después de la limpieza, se pueden crear nuevas características o modificar las existentes para mejorar el análisis o los modelos de machine learning. Incluye la creación de características combinando otras, la generación de características polinómicas, la normalización y la codificación de variables categóricas.

Buenas Prácticas y Exportación de Datos

Conservar los datos brutos

Siempre guardar una copia original antes de cualquier modificación.

Exportación de Datos

Pandas permite exportar DataFrames a varios formatos utilizando métodos como to_csv(), to_excel(), to_json(), to_hdf(), to_sql() o to_pickle(). Se pueden personalizar opciones como incluir el índice en el archivo CSV (index=True por defecto en to_csv()).

Documentar el proceso

Mantener un registro de los pasos realizados.

Escribir funciones reutilizables

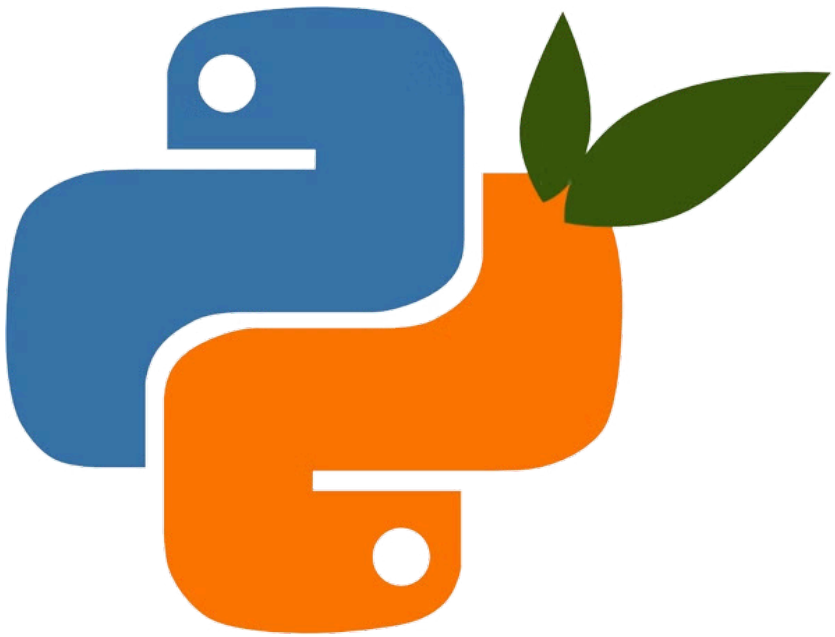
Encapsular tareas comunes de limpieza en funciones.

Utilizar listas de verificación

Para no omitir pasos importantes.

La limpieza de datos con Pandas es un proceso iterativo y esencial para garantizar la calidad y confiabilidad de cualquier análisis de datos. La clave del éxito radica en comprender los datos, identificar los problemas específicos y aplicar las herramientas de Pandas de manera efectiva y documentada. Invertir tiempo en la limpieza de datos conduce a resultados más precisos y decisiones mejor informadas.

Pasamos ha hacer una práctica...



¿Voluntari@?