Análisis de archivo para perfilamiento de datos.

Para realizar una limpieza de datos en un DataFrame con base el análisis del archivo HTML generado, se siguieron los siguientes pasos:

1. Analizar el informe HTML de perfilamiento

El informe generado dará detalles sobre los problemas en los datos, tales como:

- **Valores nulos**: Columnas con muchos valores faltantes.
- **Correlaciones**: Relación entre variables que podría indicar redundancia.
- **Tipos de datos incorrectos**: Columnas que deberían tener un tipo de dato diferente.

En el archivo HTML se revisaron las siguiente secciones:

- **"Missing Values"** (valores faltantes)
- **"Correlations"** (correlaciones altas)
- **"Data types"** (tipos de datos)

2. Limpiar los datos de acuerdo con el informe

Con base en el informe de perfilamiento, se aplicaron diferentes técnicas para limpiar el DataFrame:

- **Eliminar o imputar valores nulos**

  Si se tienen muchas filas o columnas con valores nulos, se utilizaron diferentes criterios para eliminación e imputación de valores nulos.

- **Cambiar tipos de datos**

  Si el perfilamiento indica que alguna columna tiene un tipo de dato incorrecto, se debe llevar al formato adecuado. Por ejemplo, una columna de fechas que se trata como texto, deberá ser cambiada a un formato adecuado para fechas.

- **Eliminar columnas innecesarias**

  Si hay columnas que no aportan valor al análisis, fueron eliminadas.

3. Repetir el proceso

Después de realizar las modificaciones, se genera un nuevo informe de perfilamiento para verificar que los problemas hayan sido resueltos y que los datos estén más limpios.

4. Guardar los datos limpios

Los datos limpios fueron guardados en un archvio parquet.

5. Resumen

- Generar el informe de perfilamiento con ydata_profiling.
- Analizar los puntos críticos del informe (valores nulos, correlaciones, etc).
- Limpiar los datos aplicando técnicas como imputación, cambio de tipos de datos, eliminación de columnas irrelevantes, etc.
- Guardar los datos limpios.