Curso base

INTELIGENCIA ARTIFICIAL DESDE CERO

LECCIÓN 14
Calidad y limpieza de datos

Cómo revisar, corregir y preparar una tabla sencilla antes de analizarla con apoyo de Google Colab e IA generativa

Lección	14	**Tema**	Calidad y limpieza de datos
Nombre del participante	_____	**Fecha**	_____

Actividad central

Actividad principal: revisión y limpieza de una tabla sencilla en Google Colab para detectar vacíos, corregir nombres, ordenar tipos de datos y eliminar duplicados.
Producto esperado: notebook básico con una tabla revisada, varias correcciones sencillas y un archivo limpio listo para seguir trabajando.

Meta de la lección. Comprender que antes de analizar, graficar o usar una tabla en ejercicios de inteligencia artificial conviene revisar su calidad. En esta lección el participante aprenderá, paso a paso, a detectar valores vacíos, espacios innecesarios, nombres poco claros, tipos de datos mal interpretados y registros repetidos dentro de Google Colab.

Idea central

Una tabla puede verse bien a primera vista y aun así contener problemas que dificultan el análisis. Si una fecha está incompleta, si un nombre aparece con espacios, si una columna numérica se lee como texto o si una fila está repetida, las conclusiones pueden salir mal.

Limpiar los datos no significa complicar el trabajo. Significa preparar la información para que luego sea más confiable, más clara y más fácil de interpretar.

¿Qué significa calidad y limpieza de datos?

Hablar de **calidad de datos** significa preguntarse si la información está completa, clara, coherente y útil para el propósito del ejercicio. Hablar de **limpieza de datos** significa hacer correcciones sencillas para mejorar esa información: quitar espacios, revisar vacíos, unificar formatos, corregir tipos y eliminar repeticiones innecesarias.

Problema frecuente	Ejemplo sencillo	Acción recomendable
Valor vacío	Falta una fecha o una categoría.	Detectar con `isnull()` y decidir si se corrige, se completa o se excluye.
Espacios innecesarios	Aparece `Cuaderno "` en lugar de `Cuaderno"`.	Usar `strip()` para limpiar el texto.
Formato desigual	Una categoría sale como `Papelería`, `papelería` y `PAPELERÍA`.	Unificar mayúsculas y minúsculas.
Tipo de dato incorrecto	La columna `ventas` se lee como texto y no como número.	Convertir con `to_numeric()`.
Registro duplicado	La misma fila aparece dos veces.	Eliminar duplicados con `drop_duplicates()`.

Antes de empezar en Google Colab

Para esta guía conviene trabajar con un archivo pequeño llamado `ventas_limpieza.csv`. Puede elaborarlo en Excel, Google Sheets o LibreOffice Calc y guardarlo en una carpeta fácil de encontrar. La idea es practicar con una tabla corta que tenga algunos detalles por corregir, porque así la limpieza se vuelve visible y comprensible.

fecha	producto	ventas	categoría
2026-01-10	`Cuaderno`	15	Papelería
2026-01-11	Lápiz	20	`papelería`
vacío	Borrador	12	*vacío*
2026-01-13	Regla	18	Oficina
2026-01-13	Regla	18	Oficina

¿Qué conviene observar desde el comienzo?

La tabla tiene una fecha vacía, espacios alrededor de algunos textos, diferencias en la escritura de la categoría y una fila repetida. Todo esto es suficiente para practicar una limpieza inicial sin complicar demasiado el ejercicio.

Paso a paso 1. Abrir la tabla y revisar qué problemas aparecen

Secuencia sugerida

1. Abra Google Colab y cree un notebook con un título como: `Guía 14 - Calidad y limpieza de datos`.
2. Suba el archivo `ventas_limpieza.csv` al entorno de trabajo.
3. Lea la tabla con `pandas` y observe sus primeras filas.
4. Revise la estructura general de la tabla: cuántas filas tiene, qué tipo de dato reconoce `pandas` y cuántos valores vacíos aparecen por columna.

Carga inicial del archivo

```
from google.colab import files
files.upload()

import pandas as pd

df = pd.read_csv("ventas_limpieza.csv")
df.head()

df.shape
df.info()
df.isnull().sum()
```

¿Qué debe mirar aquí?

Si una columna que debería ser numérica aparece como `object`, conviene revisarla. Si una columna muestra valores vacíos, también debe tomarse nota. Esta revisión inicial no corrige todavía: simplemente ayuda a detectar dónde vale la pena intervenir.

Prompt sugerido con IA generativa

Ya cargué `ventas_limpieza.csv` en Google Colab. Actúa como tutor para principiantes y explícame con lenguaje sencillo qué me dicen `df.shape`, `df.info()` y `df.isnull().sum()`. Después ayúdame a redactar dos observaciones cortas sobre posibles problemas de calidad en la tabla.

Paso a paso 2. Limpiar nombres de columnas y textos con espacios

En muchas tablas aparecen columnas con espacios invisibles o textos escritos de formas distintas. Eso causa confusión al consultar o agrupar información. Por eso conviene hacer una limpieza simple de nombres y columnas de texto.

Limpieza básica de nombres y textos

```
df.columns = df.columns.str.strip().str.lower()

df["producto"] = df["producto"].str.strip().str.title()
df["categoria"] = df["categoria"].fillna("Sin dato").str.strip().str.title
    ()

df.head()
```

Lectura pedagógica del cambio

Después de esta limpieza, `Cuaderno  "` pasa a `Cuaderno"`, y una categoría vacía puede quedar como `"Sin dato"`. Además, los nombres de las columnas quedan uniformes y más fáciles de usar en el código.

Prompt sugerido con IA generativa

Explícame paso a paso qué hace este bloque de limpieza: `df.columns = df.columns.str.strip().str.lower()`, `df["producto"] = df["producto"].str.strip().str.title()` *y* `df[çategoria"] = df[çategoria"].fillna("Sin dato").str.strip().str.title()`. *No uses lenguaje técnico y dame un ejemplo antes y después.*

Paso a paso 3. Corregir tipos de datos, revisar fechas y eliminar duplicados

Una tabla puede necesitar también correcciones de tipo. Las fechas conviene convertirlas a formato de fecha y los valores numéricos a formato numérico. Después, si una fila aparece repetida sin necesidad, puede eliminarse.

Conversión de tipos y eliminación de duplicados

```
df["fecha"] = pd.to_datetime(df["fecha"], errors="coerce")
df["ventas"] = pd.to_numeric(df["ventas"], errors="coerce")

df = df.drop_duplicates()

df.info()
df
```

¿Por qué `errors=çoerce"` puede ser útil?

Cuando se usa `errors=çoerce"`, `pandas` marca como vacíos los datos que no logra convertir correctamente. Esto no es un error pedagógico; al contrario, ayuda a detectar con claridad qué registros requieren atención. En este ejemplo, la fecha vacía seguirá apareciendo como faltante, pero la fila repetida dejará de contar dos veces.

Prompt sugerido con IA generativa

Tengo una columna `fecha` y una columna `ventas`. Explícame por qué conviene convertirlas con `pd.to_datetime` y `pd.to_numeric`, qué significa `errors="coerce"` y cómo puedo explicar a un principiante por qué `drop_duplicates()` mejora la tabla.

Paso a paso 4. Guardar una versión limpia para seguir trabajando

Una buena práctica consiste en no perder de vista el resultado de la limpieza. Por eso conviene guardar una nueva versión del archivo, con otro nombre, para usarla en visualización, análisis o ejercicios posteriores.

Guardar una versión limpia

```
df_limpio = df.copy()
df_limpio.to_csv("ventas_limpieza_final.csv", index=False)

df_limpio.head()
```

Resultado esperado

Así queda un archivo separado de la versión original. Este detalle es útil porque permite comparar antes y después, y ayuda a que el participante comprenda que la limpieza es una etapa del proceso y no una acción invisible.

Prompt sugerido con IA generativa

Actúa como acompañante pedagógico. Ayúdame a explicar por qué es útil guardar un archivo nuevo después de limpiar los datos y propón una frase corta que pueda escribir en mi notebook para dejar constancia de lo que corregí.

Errores frecuentes al comenzar

Situación	Posible causa	Qué conviene hacer
La columna no aparece en el código	El nombre tiene mayúsculas, tildes o espacios distintos.	Revise `df.columns` y copie el nombre exacto o normalícelo.
Los números no se pueden promediar	La columna se leyó como texto.	Conviértala con `pd.to_numeric()`.
Las fechas salen desordenadas o como texto	No se convirtieron a formato fecha.	Use `pd.to_datetime()`.
Aparecen demasiados vacíos	La tabla original tiene celdas en blanco o formatos mezclados.	Use `isnull().sum()` y revise cada caso.
Los resultados se repiten más de lo esperado	Hay filas duplicadas.	Pruebe `drop_duplicates()` y vuelva a contar.

Uso pedagógico de la IA generativa en esta guía

La IA generativa puede ayudar mucho cuando explica qué significa un valor vacío, por qué una columna numérica terminó como texto o cómo redactar observaciones sencillas sobre la limpieza realizada. También puede servir para revisar si una instrucción fue escrita con claridad o para sugerir formas simples de documentar el proceso.

Sin embargo, la tabla real sigue siendo el punto de referencia principal. La respuesta de la IA debe contrastarse siempre con lo que aparece en el notebook.

Sugerencia didáctica

Conviene pedir a la IA explicaciones cortas, paso a paso y con ejemplos antes y después. Eso ayuda a principiantes que todavía no dominan el vocabulario técnico del trabajo con datos.

Actividad principal

Cada participante trabajará con una tabla pequeña que contenga algunos problemas sencillos de calidad. Primero la cargará en Google Colab. Después revisará su estructura, detectará vacíos, limpiará textos, corregirá tipos básicos y eliminará duplicados. Finalmente guardará una versión limpia y escribirá un breve comentario sobre lo que cambió.

Producto esperado

Un notebook corto, claro y comentado, con evidencia visible del antes y del después de la limpieza.

Lista rápida para revisar el notebook

Revisión	Sí / No
¿El notebook tiene un título claro y una breve explicación del objetivo?	_____
¿El archivo fue cargado correctamente y se visualizaron las primeras filas?	_____
¿Se revisaron `shape`, `info` e `isnull().sum()`?	_____
¿Se limpiaron nombres de columnas o textos con espacios?	_____
¿Se corrigió al menos un tipo de dato importante?	_____
¿Se revisaron o eliminaron duplicados?	_____
¿Se guardó una versión limpia del archivo?	_____
¿Se usó la IA como apoyo, pero verificando siempre con la tabla real?	_____

Mini evaluación

Responda con sus palabras

1. Explique por qué no conviene analizar una tabla sin revisarla primero.

2. Mencione dos problemas sencillos de calidad de datos que puedan aparecer en un archivo CSV o Excel.

3. Describa una diferencia entre detectar un problema y corregirlo.

4. Escriba un prompt claro para pedir a la IA una explicación sencilla sobre valores vacíos o duplicados.

Cierre

Limpiar datos es una habilidad fundamental porque mejora la confianza en lo que después se grafica, se resume o se utiliza en ejercicios más avanzados. En esta lección no se busca un tratamiento técnico complejo. Lo importante es comprender una secuencia simple y poderosa: cargar, revisar, detectar, corregir, guardar y documentar. Esa rutina prepara muy bien el camino para seguir trabajando con datos de manera responsable.

Tarea para practicar

Tome uno de los archivos sencillos de las lecciones anteriores, por ejemplo `ventas.csv`, `estudiantes.xlsx` o `inventario.csv`, y revíselo con la misma lógica de esta guía. Busque si hay vacíos, diferencias en la escritura de categorías, columnas que deban convertirse o registros repetidos. Luego redacte tres líneas cortas explicando qué encontró y qué decidió corregir.

También puede pedir a la IA que le proponga una lista de revisión para tablas pequeñas y compararla con la que usted ya utilizó. El objetivo no es pedir una solución automática, sino fortalecer el criterio para reconocer cuándo una tabla está lista para el siguiente paso del curso.