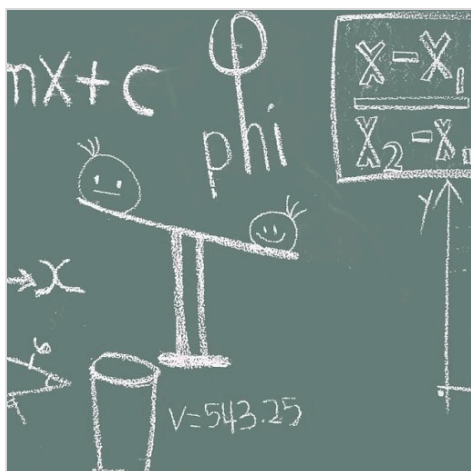


L U M B R E



Limpieza de datos: el 80% del trabajo que nadie te muestra

Grado: 6

Materia: Pensamiento Matemático VI

03/10/2026

Antes de empezar

Tabla de cinco problemas; outlier $\neq$ error; tres estrategias de faltantes; bitácora de limpieza.

QUÉ APRENDERÁS

- Diagnosticar y resolver los cinco problemas típicos de datos (faltantes, duplicados, erróneos, inconsistentes, outliers), documentar la limpieza y elegir imputación según la distribución.

Los datos crudos casi nunca sirven

El dato que llega es siempre sospechoso

Lo que se ve en publicaciones y dashboards es el dato limpio. Lo que nadie cuenta es que el 60-80% del tiempo de un analista se va en limpiar: valores faltantes, duplicados, erróneos de captura, inconsistencias de formato. Aprender a limpiar es aprender a no confiar ciegamente en lo que parece claro.

Los cinco problemas más comunes



Imagen libre de Wikimedia Commons (CC BY 2.0, autor: Cristina Cifuentes)

Problema	Cómo se ve	Qué hacer
faltante	celdas vacías, NA, espacios	decidir si se elimina o se imputa
duplicado	dos filas con el mismo id	quedarse con una (criterio explicitado)
erróneo	edad 250, nota 15 sobre 10	regresar a la fuente; no inventar
inconsistente	México, mex., MÉXICO	normalizar a un formato
outlier	un ingreso 10 veces mayor que el resto	verificar antes de quitar

Una decisión clave: los outliers NO son errores por definición. Un dato atípico puede ser perfectamente válido (el ingreso de una empresa enorme mezclado con mipymes). Quitar un outlier sin justificación es manipular el resumen; dejarlo sin contextualizar es distorsionar la media.

Valores faltantes: tres estrategias honestas

1. ELIMINAR fila: si el faltante es aleatorio y no cambia n.
2. IMPUTAR con mediana: si la variable es numérica y la distribución es asimétrica.
3. IMPUTAR con moda: si es categórica.
4. DECLARAR: reportar cuántos faltan y con qué criterio se resolvió.
5. ENCUESTA DE SEGUIMIENTO: si el faltante no es aleatorio (mucha gente que no quiso responder), eso es el hallazgo.

El 5% que no era ruido

Ejemplo resuelto: una encuesta de transporte tenía 5% de campos vacíos en ingreso. Al analizar, se vio que los vacíos concentraban en el decil más alto (la gente rica no declara). Imputar con mediana habría distorsionado. Se optó por reportar dos series: con y sin no-declarantes, y el fenómeno (evasión de ingreso) se volvió parte del resultado.

La bitácora de limpieza (lo que separa lo profesional de lo casero)

- Qué se quitó y por qué (con conteo antes/después).
- Qué regla de imputación se aplicó.
- Qué variables se crearon, recodificaron o combinaron.
- Qué filas se excluyeron del análisis final (para reproducibilidad).

1. Un outlier es...

- ☐ siempre un error de captura
- ☐ un valor atípico que puede ser válido o erróneo
- ☐ lo mismo que valor faltante
- ☐ un duplicado

2. Imputar siempre con el promedio es la mejor estrategia.

- ☐ Verdadero
- ☐ Falso

3. Abre cualquier tabla tuya y cuenta cuántos huecos, duplicados y formatos distintos hay por columna.

El diagnóstico sorprende

4. El 60-80% del tiempo se va en _____ ; los outliers se _____ antes de quitar; los faltantes se reportan con la regla de imputación _____ .

ACTIVIDAD INTERACTIVA DISPONIBLE EN
LUMBRE

5. ¿Documentarías tu
limpieza?



VIDEO

28) Limpieza de datos con pandas: nulos, duplicados y tipos en 4 líneas



Limpieza de datos: valores faltantes, duplicados y outliers.



Escanea para verlo en Lumbre

<https://www.youtube.com/watch?v=X8x0JBgoe8Q>

Para mirarlo de otra manera

La regla del tres antes de tocar

Antes de limpiar una celda, se abre la tabla tres veces: (1) ver la columna completa, (2) ordenar de mayor a menor, (3) filtrar por el valor dudoso. Casi todo hallazgo aparece en una de esas tres revisiones.

6. ¿Qué columna sospechosa de tu tabla ordenarías primero?

La que tiene el máximo raro

Referencia curricular

Grado:	6
Materia:	Pensamiento Matemático VI

Versión interactiva completa (con actividades, videos y autoevaluación):



<https://lumbre.lat/object/pmat6-u1-cap-limpiar>

Material educativo generado desde Lumbre · lumbre.lat