

L U M B R E



Valores faltantes y estrategias de imputación

Grado: 3

Materia: Preparación y Calidad de Datos · Análisis de Datos

11/10/2026

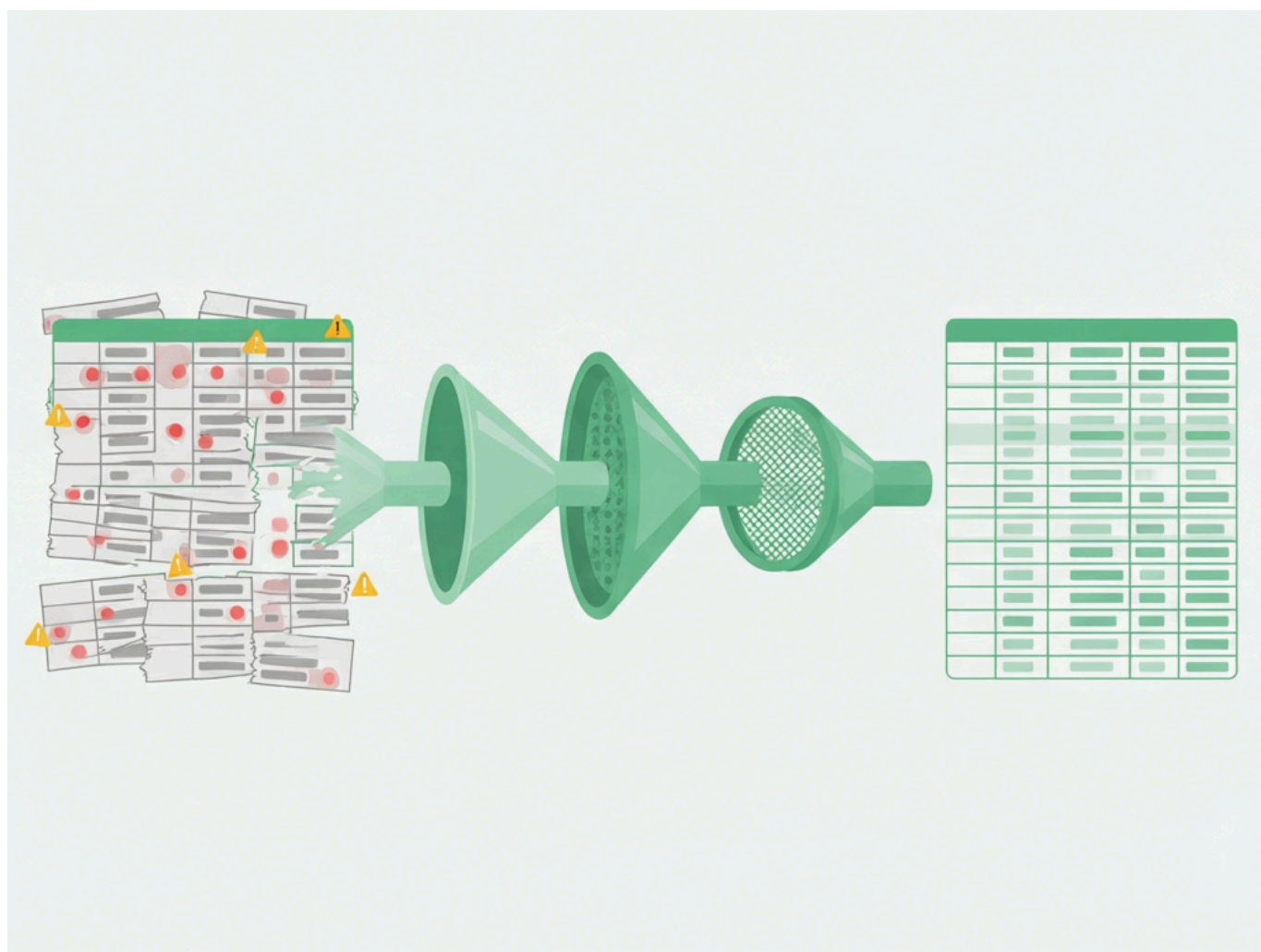
Antes de empezar

Casi ningún conjunto de datos real está completo. Antes de rellenar huecos a ciegas hay que preguntarse por qué faltan los valores, porque el mecanismo que los genera decide qué tratamiento es válido y cuál introduce un sesgo silencioso.

QUÉ APRENDERÁS

- Diagnosticar por qué faltan datos (mecanismo de los ausentes), cuantificar su impacto y elegir la estrategia de tratamiento adecuada: eliminar, imputar o modelar.

Valores faltantes y estrategias de imputación



Preparacion y calidad de datos: faltantes, atipicos, codificacion y pipelines.

Mapa conceptual

- **Calidad de Datos**

- Faltantes
 - MCAR
 - MAR
 - Imputación
- Atípicos
 - IQR
 - Z-score
- Codificación
 - One-hot
 - Label
 - Escalado
- Deduplicación
 - Similitud
 - Fuzzy
- Pipelines
 - Reproducibilidad
 - Versionado

El mapa mental interactivo es interactivo en la versión online.



Por qué falta importa más que cuánto falta

Un dato que falta no es un cero ni un despiste: puede faltar por azar, por conveniencia o precisamente por el valor que habría tenido. Tratar todos los huecos igual es la receta para inventar conclusiones.

El problema no es el hueco, es el motivo

Dos conjuntos con un diez por ciento de valores ausentes pueden exigir tratamientos opuestos según por qué faltan. La diferencia entre un cero ausente y un no-sabe-no-contestó cambia por completo la forma correcta de proceder.

MCAR: falta al azar

Missing Completely At Random: la probabilidad de que falte un valor no depende de nada, ni de las demás variables ni del propio valor ausente. Es el único caso en que

eliminar los registros incompletos no introduce sesgo, solo pierde precisión. En la práctica es raro.

MAR: falta según lo observado

Missing At Random: la ausencia depende de otras variables que sí observamos, pero no del valor que falta. Por ejemplo, los hombres responden menos a una encuesta de salud, y eso se explica por el sexo, que conocemos. Imputar condicionando a las observables recupera la validez.

MNAR: falta por el propio valor

Missing Not At Random: lo que falta depende del propio valor ausente. Las personas con rentas altas ocultan más su ingreso. Este es el caso terrible: ningún método basado solo en los datos observados puede corregirlo sin suposiciones externas, y todo tratamiento puede sesgar.

Diagnosticar el mecanismo

Se comparan los registros completos con los incompletos en las variables observadas, y se estudia si la ausencia se correlaciona con ellas. No se puede probar MNAR con los datos (el valor clave no está), pero sí descartar MCAR y sospechar de MAR cuando la presencia depende de covariables.

Eliminar: cuándo es honesto

Borrar filas o columnas con huecos solo es defendible bajo MCAR y con pérdida de datos pequeña. Bajo MAR o MNAR, eliminar condiciona la muestra a quien respondió y sesga sistemáticamente las estimaciones. El borrado silencioso de filas es la forma más común de error de datos.

Eliminación por lista

Descartar cualquier fila con un solo hueco (complete case) puede destruir muestra: con muchas variables y pocos huecos cada una, la intersección completa queda minúscula y posiblemente sesgada. Conviene solo para columnas casi completas o análisis exploratorios.

Imputar con la media o mediana

Rellenar con el valor central es simple pero empobrecedor: aplasta la variabilidad, crea un pico artificial y sesga las relaciones entre variables al tirar de todos los huecos hacia el centro. Útil solo como referencia ingenua, nunca como decisión final.

Imputación por regresión

Ajustar un modelo que prediga la variable ausente a partir de las observadas y rellenar con esa predicción. Mejora a la media porque usa información, pero ignora el error de predicción, así que aún subestima la varianza. Es un paso intermedio.

Imputación múltiple

En lugar de un relleno, se generan varios conjuntos completos añadiendo incertidumbre a cada imputación, se analiza cada uno y se combinan los resultados (reglas de Rubin). Recupera la varianza perdida y propaga la incertidumbre del hueco al resultado final: el estándar cuando MAR es plausible.

Método de encadenadas

La imputación múltiple por ecuaciones encadenadas (MICE) simula imputando cada variable con huecos desde un modelo de las demás, varias pasadas hasta estabilizar. Flexible para mezclar tipos de variables y el enfoque más extendido en la práctica estadística.

Imputar y añadir indicador

Añadir una columna indicadora de si el valor faltaba permite al modelo aprender un efecto propio de la ausencia. Ojo: con una variable completamente imputada, el indicador puede volverse linealmente dependiente; se combina bien con imputación por modelo.

K-vecinos y métodos de vecinos

Rellenar con la media de los registros más parecidos (k vecinos más próximos) capta estructura local sin un modelo paramétrico. Sensible a la métrica de distancia y a la escala de las variables, que hay que estandarizar antes.

Cuidado con filtrar antes de imputar

La imputación debe respetar la fuga de información: si se imputa usando toda la tabla, incluido el futuro conjunto de validación, se contamina el modelo. En flujos de aprendizaje automático, la imputación se ajusta solo con los datos de entrenamiento.

Errores frecuentes

- Rellenar con cero un hueco, mezclando ausencia con un valor real de magnitud nula.

- Imputar con la media y luego celebrar que la correlación «apenas cambió», cuando la imputación la creó.
- Borrar filas incompletas sin comprobar si la ausencia estaba relacionada con la respuesta.

1. Bajo un mecanismo MNAR, la estrategia honesta es:

- ☐ Eliminar siempre las filas incompletas
- ☐ Imputar por la media
- ☐ Requiere modelar explícitamente la falta o sensibilizar, pues ningún método solo con lo observado corrige el sesgo
- ☐ Rellenar con ceros

2. Imputar los huecos con la media de la variable deja intacta su variabilidad.

- ☐ Verdadero
- ☐ Falso

Reportar la falta

Toda imputación debe documentarse: cuánto faltaba, bajo qué supuesto y con qué método. Un análisis que no declara sus datos faltantes es imposible de auditar, y los resultados pueden depender más del relleno que de la evidencia.

Imputación en series temporales

Con datos ordenados en el tiempo, rellenar con el vecino anterior o interpolar respeta la autocorrelación. Hay que vigilar huecos largos y no mirar hacia el futuro al imputar un punto, para no romper la causalidad de la serie.

ACTIVIDAD INTERACTIVA DISPONIBLE EN
LUMBRE

3. Mecanismos y remedios



4. Ordena las opciones de menor a mayor sofisticación para un hueco:

Elementos:

- eliminar filas incompletas
- rellenar con el valor central
- predecir con otras variables
- imputar con incertidumbre y combinar

Borrar	. . .
Media	. . .
Regresion	. . .
Multiple	. . .

VIDEO

Video complementario

Recurso audiovisual para reforzar los conceptos.



Escanea para verlo en Lumbre

<https://www.youtube.com/watch?v=rEwBnM0qSf0>

5. En una encuesta de ingresos, las rentas más altas se declaran menos. ¿Qué mecanismo de falta es y por qué imputar por la media sería especialmente engañoso aquí?

Documentación pandas <https://pandas.pydata.org/docs/Guía de referencia para limpieza de datos>.

Referencia curricular

Grado:	3
Materia:	Preparación y Calidad de Datos · Análisis de Datos

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



<https://univ.lumbre.lat/object/cdia-pcal-001-valores-faltantes-y-estrategias-de-imputacion>

Material educativo generado desde Lumbre · lumbre.lat