

L U M B R E



Atípicos y consistencia

Grado: 3

Materia: Preparación y Calidad de Datos · Análisis de Datos

11/10/2026

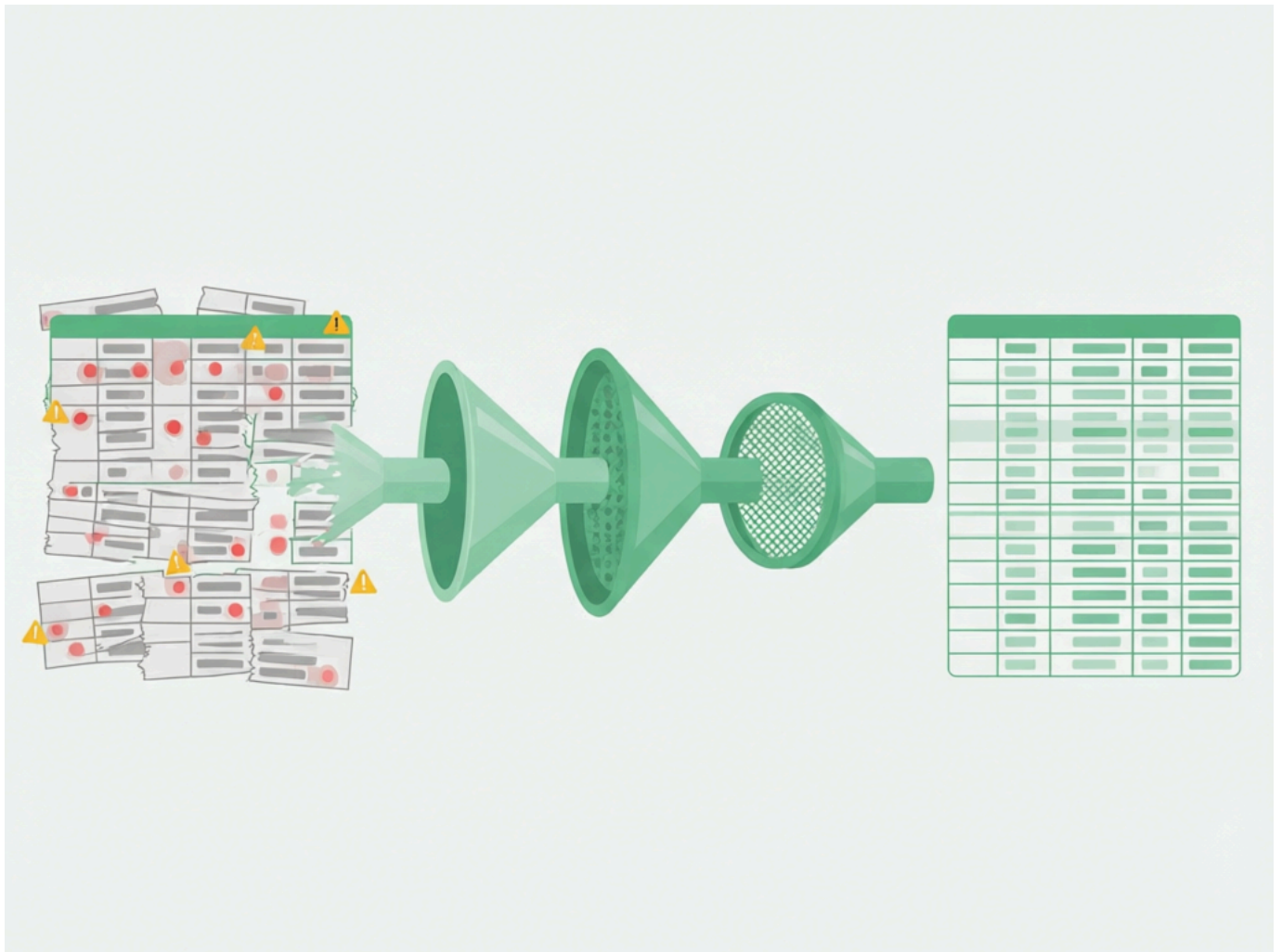
Antes de empezar

Un número absurdo delata casi siempre un error de captura; un valor extremo perfectamente posible puede ser el hallazgo más importante del estudio. Esta lección trata de distinguir uno del otro y de verificar que los datos se contradicen a sí mismos.

QUÉ APRENDERÁS

- Detectar valores atípicos y violaciones de reglas de consistencia, distinguiendo error de señal real, y decidir el tratamiento sin destruir información válida.

Atípicos y consistencia



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.



Error o hallazgo

Un valor atípico no es un enemigo a borrar: a veces es un dedo equivocado en el teclado, y a veces la tormenta que estabas buscando. Detectar es fácil; decidir qué hacer con él, difícil.

Qué es un atípico

Un valor que se desvía tanto del resto que parece generado por un proceso distinto. La definición es relativa a la distribución esperada: un ingreso de dos millones es atípico para una encuesta de hogares y corriente en una cartera de fondos.

Tres orígenes

- Error: dedo torpe, unidad equivocada, desbordamiento de un campo, un signo al revés.
- Variabilidad natural: el extremo legítimo de una distribución de cola pesada.
- Nueva física: un fenómeno real aún no visto, que el modelo «no espera».

Detectar con estadísticos

Reglas como tres desviaciones típicas, o el rango intercuartílico multiplicado por un factor, marcan candidatos. Son filtros de atención, no sentencias: señalan dónde mirar y cuánta observación merece un valor, no prueban que sea erróneo.

La trampa de la regla tres sigma

Bajo normalidad, la regla tres desvíos marca unos pocos por mil, pero los datos reales rara vez son normales. En distribuciones de cola pesada o sesgadas, un umbral ingenuo tacha de atípicos valores plenamente esperables. Conocer la forma importa.

Distancia de Mahalanobis

Para detectar anomalías multivariantes, la distancia de Mahalanobis mide cuánto se aleja un punto del centro teniendo en cuenta la covarianza: atrapa combinaciones extrañas de variables que vistas de una en una parecen normales.

Reglas de consistencia

Más allá de lo extremo, los datos pueden contradecirse: una fecha de fin anterior al inicio, una edad de dos años con estado civil casado, un código postal que no corresponde a la provincia. Estas violaciones lógicas son errores seguros, y se detectan con reglas derivadas del dominio.

Dominio manda

Qué es imposible lo decide el conocimiento del negocio, no la estadística. Un peso corporal de cinco kilos en un adulto es un error; en un recién nacido es normal. Codificar validaciones de rango y formato según el dominio es la primera línea de defensa.

Rangos y formatos

Verificar que cada campo cae entre sus límites físicos y lógicos, y que respeta el formato: fechas parseables, identificadores con la máscara correcta, campos obligatorios presentes. Barato de implementar y captura una fracción enorme de errores.

No borrar a lo bruto

Eliminar todo atípico por sistema destruye señal y sesga la distribución hacia el centro. Primero hay que investigar el origen: si es error verificable, corregirlo o vaciarlo; si es real, conservarlo y asegurarse de que el modelo lo soporte.

Winsorizar y transformar

Reemplazar los extremos por percentiles (winsorizar) o aplicar transformaciones (logaritmo) atenúa la palanca de los atípicos sin eliminarlos del todo. Útil para métodos sensibles a colas, pero cambia la escala y hay que reportarlo.

Robustez frente a exclusión

Muchas veces la respuesta correcta no es tocar los datos sino elegir métodos robustos: medianas en lugar de medias, regresión robusta, métricas resistentes. Dejar el atípico y usar un estimador que no le haga caso suele ser más honesto que borrarlo.

Atípicos en modelos predictivos

Un punto muy influyente puede sesgar un modelo sensible (mínimos cuadrados) y ser ignorado por otro (árboles). Medir influencia —distancia de Cook, leverages— dice cuánto cambiaría el ajuste sin él, que es la pregunta relevante, no si se ve raro.

Fuga de la detección

Detectar atípicos o calcular umbrales usando el conjunto completo (incluida validación) filtra información. Las reglas de limpieza y los límites deben estimarse solo con entrenamiento y aplicarse después, igual que la imputación.

Errores frecuentes

- Borrar todo lo que se aleja de la media por principio, achatando colas reales.
- Confundir un error de unidad (metros por kilómetros) con un fenómeno interesante.
- No escribir reglas de consistencia y dejar que las contradicciones lógicas pasen al modelo.

1. Un valor extremo detectado por una regla estadística debe, ante todo:

- ☐ Eliminarsen inmediatamente
- ☐ Investigarse su origen antes de decidir
- ☐ Reemplazarse por la media
- ☐ Considerarse automáticamente un fraude

2. Las reglas de consistencia se derivan del conocimiento del dominio, no solo de la distribución de los datos.

- ☐ Verdadero
- ☐ Falso

Documentar cada intervención

Cada valor corregido, vaciado o recortado debe quedar anotado con su motivo. La reproducibilidad y la auditoría exigen poder reconstruir qué se tocó y por qué; un conjunto de datos «limpio» sin bitácora es una caja negra.

Detección no supervisada

Cuando no hay etiqueta de error, métodos de detección de anomalías (bosques de aislamiento, vecindad local, reconstrucción por autoencoder) puntúan cuán inusual es cada registro, priorizando la revisión humana en los más raros.

ACTIVIDAD INTERACTIVA DISPONIBLE EN
LUMBRE

3. Atípicos y consistencia



4. Ordena el manejo responsable de un valor extremo:

Elementos:

- marcar candidatos con una regla
- investigar si es error o señal
- corregir, conservar o usar método robusto
- anotar qué se hizo y por qué

Detectar	. . .
Origen	. . .
Decidir	. . .
Documentar	. . .

Interpretable ML <https://christophm.github.io/interpretable-ml-book/> Influencia de observaciones y detección de anomalías explicada con claridad.

VIDEO

Video complementario

Recurso audiovisual para reforzar los conceptos.



Escanea para verlo en Lumbre

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

5. Tu dataset de ventas marca un pedido de importe un millón de veces mayor que el resto, con una fecha válida y cliente conocido. ¿Qué pasos seguirías antes de decidir borrarlo?

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-002-atipicos-y-consistencia>

Material educativo generado desde Lumbre · lumbre.lat