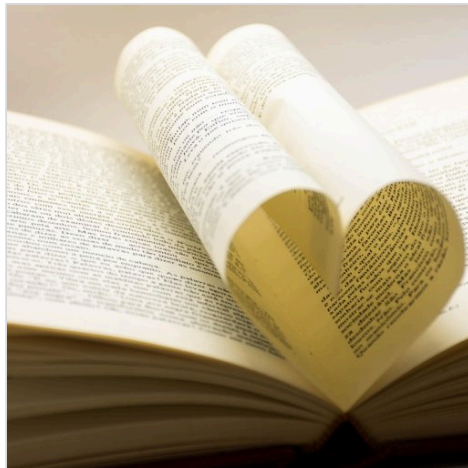


L U M B R E



Univariante: distribución y atípicos

Grado: 4

Materia: Análisis Exploratorio de Datos · Análisis de Datos

11/10/2026

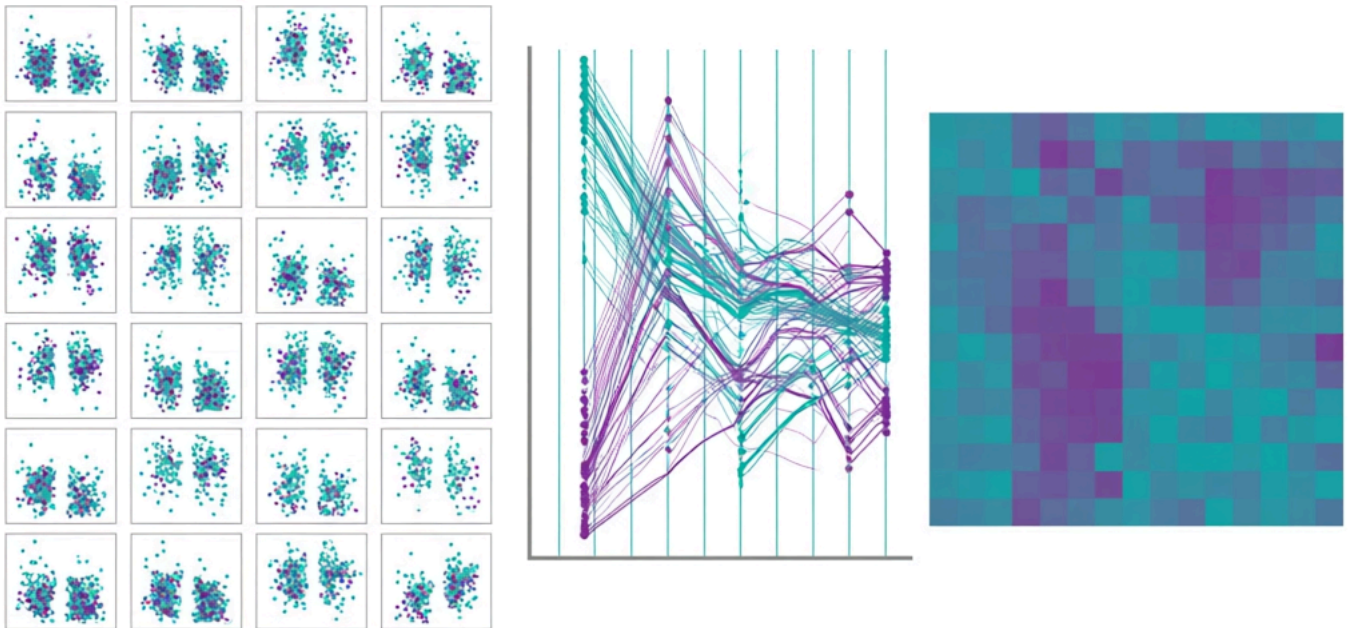
Antes de empezar

La exploración empieza por lo simple: una variable cada vez. Entender a fondo cómo se distribuye cada columna, dónde se concentra y dónde se desmanda, construye el vocabulario con el que después se leerán las relaciones.

QUÉ APRENDERÁS

- Caracterizar una variable a la vez: su tipo, forma, centro y dispersión, y detectar valores atípicos e imposibles como fundamento de la exploración.

Univariante: distribución y atípicos



Análisis exploratorio de datos: visualización multivariante y detección de patrones.

Mapa conceptual

- **EDA**
 - Univariante
 - Histogramas

- Boxplots
- Atípicos
- Bivariante
 - Scatter
 - Correlacion
 - Confusion
- Multivariante
 - Pairs plot
 - Parallel coord
- Modelo
 - Hipotesis
 - Selecccion

El mapa mental interactivo es interactivo en la versión online.



Una cosa a la vez

Antes de preguntar cómo se relacionan dos cosas, hay que saber qué es cada una. Estudiar una variable a la vez, ver su forma completa, es el albañil poniendo bien el primer ladrillo.

Empezar por el tipo

La estrategia univariante depende del tipo. Una numérica continua admite histograma y cuantiles; una discreta, un diagrama de barras; una categórica, un conteo de frecuencias por categoría. Aplicar la media a una variable nominal es un sin sentido que el software permite pero la razón no.

La forma de la distribución

¿Un pico o dos? ¿Simétrica o sesgada? ¿Colas cortas o pesadas? La forma lo dice casi todo sobre una variable: condiciona qué resumen es adecuado, qué transformación convendría y qué modelo sería plausible. Mirar la forma antes que el número.

Histograma y su lectura

El histograma agrupa valores en intervalos y muestra su frecuencia. Es la radiografía de la distribución. Pero miente según el ancho de intervalo elegido: demasiado ancho oculta estructura, demasiado fino exagera el ruido. Probar varios anchos es parte de mirarla bien.

Diagrama de caja

El boxplot resume con la mediana, los cuartiles y unos bigotes hasta los valores no-atípicos, marcando aparte los puntos extremos. Compara distribuciones de un vistazo y hace visible la asimetría y los atípicos sin depender de un ancho de bins.

Centro: media frente a mediana

La media usa todos los valores pero la arrastran los extremos; la mediana resiste a los atípicos pero ignora sus magnitudes. En una distribución simétrica coinciden; en una sesgada se separan, y esa separación es información valiosa sobre la asimetría.

Dispersión

El centro sin la dispersión es engañoso. Dos grupos con media cien, uno apretado y otro desbordado, son historias distintas. Desviación típica, varianza y rango intercuartílico cuantan cuánto se desparraman los valores alrededor del centro.

Sesgo y curtosis

El sesgo mide la asimetría: colas más largas a un lado. La curtosis mide el grosor de las colas respecto a la normal. Ambos resumen la forma, pero son complementos: dos variables pueden tener igual sesgo y colas muy distintas.

Detectar atípicos

Un valor atípico se desvía tanto que parece de otro proceso. La regla del rango intercuartílico (alejarse más de una vez y media el ancho intercuartílico del primer o tercer cuartil) marca candidatos, que luego hay que investigar: ¿error o realidad?

Imposibles frente a extremos

Un atípico puede ser un error flagrante (edad 250, temperatura 500) o un extremo legítimo (el cliente que sí compró un millón). Los imposibles violan el dominio y se corrigen; los extremos válidos se estudian. Distinguirlos exige conocer el fenómeno.

La cola larga

Muchos fenómenos —ingresos, tamaños de ciudad, popularidad— tienen distribuciones de cola pesada: pocos valores enormes dominan. En ellas la media es poco representativa y los atípicos abundan: no son anomalías sino la esencia del reparto.

Transformar la variable

Un logaritmo puede domeñar un sesgo enorme y hacer legible una cola larga, estabilizando la varianza. Transformar no es cosmética: cambia qué relaciones se vuelven lineales y qué resúmenes tienen sentido, y hay que elegirlo pensando en el fenómeno.

Variables categóricas

Para una categoría, el resumen es su tabla de frecuencias y la moda. Conviene ver qué tan desequilibrada está: una categoría que concentra el noventa y nueve por ciento de los casos aporta poca información y puede requerir reagrupamiento.

Ceros y vacíos confundidos

Un cero puede significar «nada» o «no medido». En una columna de ingresos, ¿cero es un pobre o un no-respondió? La exploración univariante debe desenmascarar esta ambigüedad, fuente de sesgos silenciosos cuando se mezclan ambos significados.

Errores frecuentes

- Reportar solo la media, sin dispersión ni forma, y dar la variable por «conocida».
- Tratar todo atípico como error a borrar, eliminando cola real.
- Aplicar resúmenes pensados para lo simétrico a variables muy sesgadas.

1. Si en una variable la media supera claramente a la mediana, la distribución es típicamente:

- ☐ Simétrica
- ☐ Sesgada hacia la derecha, con cola larga de valores altos
- ☐ Uniforme
- ☐ Bimodal

2. El rango intercuartílico, a diferencia del rango total, no lo altera un único valor extremo.

- ☐ Verdadero
- ☐ Falso

Contar antes de asumir

Antes de decidir que un valor raro es un error, conviene comprobar si se repite o es único, si coincide con un patrón (un solo día, un solo usuario) o aparece al azar. El contexto del raro decide su tratamiento.

Perfil de cada columna

El resultado ideal de la exploración univariante es un «perfil» por variable: tipo, forma, centro, dispersión, faltantes, atípicos y posibles problemas de dominio. Ese perfil es la ficha médica de los datos sobre la que se apoya todo lo demás.

ACTIVIDAD INTERACTIVA DISPONIBLE EN LUMBRE

3. Univariante



4. Ordena el estudio de una variable numérica:

Elementos:

- confirma que es numérica y su dominio
- mira su distribución y asimetría
- resume centro y dispersión
- localiza atípicos e imposibles

Tipo	. . .
Forma	. . .
Centro	. . .
Extremos	. . .

Distribution shapes <https://www.youtube.com/watch?v=rw43t7RCPKk> Intuición visual sobre formas de distribución, sesgo y colas.

VIDEO

Video complementario

Recurso audiovisual para reforzar los conceptos.



Escanea para verlo en Lumbre

<https://www.youtube.com/watch?v=M7tX2MJGEY>

5. Una columna «días_para_entregar» tiene media 4, mediana 2 y un máximo de 900. ¿Qué sospechas, qué comprobarías y qué harías con esos 900?

Referencia curricular

Grado:	4
Materia:	Análisis Exploratorio de Datos · Análisis de Datos

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



<https://univ.lumbre.lat/object/cdia-anex-002-univariante-distribucion-y-atipicos>

Material educativo generado desde Lumbre · lumbre.lat